Jörn Sebastian Hahn

Steuerungswirkungen zentraler Vergleichsarbeiten auf den vorgelagerten Unterricht

Testcoaching am Beispiel von Lernstand8

Jörn Sebastian Hahn

Steuerungswirkungen zentraler Vergleichsarbeiten auf den vorgelagerten Unterricht

Testcoaching am Beispiel von Lernstand8

Diese Arbeit wurde 2014 als Dissertation zur Erlangung des akademischen Grades Dr. phil. von der Fakultät für Bildungswissenschaften der Universität Duisburg-Essen angenommen (Gutachter: Prof. Dr. Isabell van Ackeren, Prof. Dr. Wilfried Bos). Die Disputation erfolgte am 13. Oktober 2014. Das Erstellen dieser Dissertation wurde durch die Friedrich-Ebert-Stiftung mit einem Stipendium gefördert.

Für Rainer Peek und das Zebra

Abstract

Following from not satisfactory results of German pupils in large-scale-assessment like TIMSS1995 or PISA2000, national standardized assessments were implemented in all sixteen German states. This tool, as a part of the current school policy called Neue Steuerung, is due to allow both support and accountability. Teachers should use the data from this standardized assessment to improve their lessons by realizing and reflecting the results single handed, as described by the Rahmenmodell der pädagogischen Nutzung von Vergleichsarbeiten by Helmke/Hosenfeld (2005). Several studies analyzed the teachers' usage of data from national standardized assessments according to this theoretical model, but none of them considered the lessons before the assessments. This is ever more astonishing, as teaching to the test effects as a result of accountability is well documented in others countries.

In this doctoral thesis, testcoaching is investigated as one of the main effects of standardized assessment (VERA8). But also chances in teaching are one of the many intended effects of the implementation of national standardized assessment in Germany. Testcoaching is analyzed from two perspectives: On the one hand, teaching to the test can limit the expressiveness of standardized assessment results, as result validity deteriorates. Particular those teachers should teach to the test, who fear bad results and who are expected to use data to improve their lessons. Under this approach, testcoaching is seen as a part of a feedback process. Testcoaching behavior is hence analyzed in connection with the Feedback Intervention Theory of Kluger/DeNisi (1996) and the Rahmenmodell der pädagogischen Nutzung von Vergleichsarbeiten by Helmke/Hosenfeld (2005). On the other hand, testcoaching can be seen as a specific lesson type. From this point of view, teachers are teaching to the test because they assume it was committed. To this the quality of teachers' lesson competence can be described by the Lehrer-Handlungskompetenz-Modell by Baumert/Kunter (2006). It is expected that teachers prepare better to national standardized assessment which have more self-efficacy and other personnel resources as other teachers do.

Two parallel survey studies were used to review the testcoaching as an effect of standardized assessment. In both studies the preparing of 8th grade maths teachers of a half of all Gymnasien in North Rhine-Westphalia were reviewed by a survey in the year 2010. I use latent class analysis to group teachers by their personnel resources (study A) and their usage of data of the standardized assessments (study B) to explore the quality of their testcoaching. In addition, the degree and the quality in general of testcoaching to the standardized assessment VERA8 in North Rhine-Westphalia is reported.

Even though all effect sizes are medium at best, both approaches can explain differences in the degree and quality of testcoaching. In fact, teachers with higher personnel resources prepare with higher quality (study A). Teachers which showed an intensive usage of data from standardized assessment VERA8 and fear bad results from the tests teach more

extensive in time and variability (study B). Differences aside, almost all surveyed teachers teach for the test in one way or the other. For example, on average, teachers prepare their students for two to three weeks and about 90% of them use old test questions in some shape or form. The results suggest revision work to the tool VERA8.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Vorwort

Den Prozess der hier vorliegenden Dissertation haben viele Menschen unterstützt. Manche von ihnen mehr am Anfang, manche mehr am Ende, viele die ganze Zeit über. Ihnen allen gebührt mein herzlicher Dank, diesen Prozess nun abschließen zu können. Eine Person, ohne die ich diesen Weg in dieser Form nicht hätte gehen können, war Rainer Peek. Von ihm stammt trotz unserer leider nur sehr kurzen Zusammenarbeit in den Wochen vor seinem unerwarteten Tod die Ausgangskonzeption dieser Arbeit. Die wenigen Treffen mit ihm reichten aus, dass Rainer Peek als Wissenschaftler mit seiner aufgeschlossenen Art ein Vorbild für mich wurde.

Ich danke Isabell van Ackeren und Wilfried Bos für die Betreuung und Unterstützung meiner Arbeit. Isabell van Ackeren hat mit großer Gelassenheit meine Besonderheiten ertragen, in die richtigen Bahnen gelenkt oder als Anlass zum Schmunzeln genommen. Sie ermöglichte mir trotz meiner formal externen Rolle vollständig in der Arbeitsgruppe integriert zu sein und auch abseits meiner Arbeit Forschungsideen umzusetzen. Sie half mir stets mit wichtigen (fachlichen) Hinweisen im Sinne des großen Ganzen und genauso aus Sicht des Kleinen und war sich auch für das Korrekturlesen nicht zuschade. Wilfried Bos war für mich der Startpunkt in die Bildungsforschung. Schon bevor ich mich für eine Promotion entschied, gab er mir die Gelegenheit Wesentliches in diesem Bereich auszuprobieren und zu lernen. Während des Prozesses unterstützte er mich dort, wo es nötig war und von mir gewünscht wurde.

Weiter danke ich den Kolleginnen und Kollegen der Arbeitsgruppe Bildungsforschung (Universität Duisburg-Essen) und des Institut für Schulentwicklungsforschung (TU Dortmund) für die Unterstützung beim Entwickeln des Erhebungsinstruments und für das wiederholte Feedback zu Teilen meiner Arbeit sowie den gemeinsamen Mittagessen. In Dortmund haben mir besonders Wolfram Rollet und Nils Berkemeyer oft geholfen. Ausdrücklich nennen möchte ich auch Martin Bonsen, der, obwohl er zeitweise an einer dauerhaften Existenz von Lernstand8 zu zweifelen schien, mich thematisch in die richtige Richtung geschoben hat und mir während meiner Zeit am IFS außerdem immer wieder mit Ratschlägen auch über das Fachliche hinaus weitergeholfen hat.

Aus Essen möchte ich besonders Denise Demski, Susanne Farwick und Kathrin Racherbäumer für das ausführliche Korrekturlesen sowie den (damaligen) studentischen Hilfskräften Stefanie Bange, Christina Funke, Alexandra Haentjes, Christian Kosmalla, Daniela Langolf und Stephan Otto für ihre Unterstützung beim Versenden der 2500 Fragebögen und die netten Gespräche zur Ablenkung hervorheben.

Wichtig war für mich auch der Austausch mit Kolleginnen und Kollegen der VERA3-Arbeitsgruppe der Universität Koblenz-Landau (vor allem Jana Große Ophoff) und Diskussionen mit Tobias Diemer und Andreas Müller.

Für die Unterstützung in statistischen Fragen und beim Einscannen der ausgefüllten Fragebögen danke ich Jens Schulze, ohne dessen Programmierkenntnisse ich auch den Umgang mit R nicht bewältigt hätte.

Ebenfalls für das ausführliche Korrekturlesen danke ich meiner Lehrerkollegin Heike Sankowsky. Meinen Mitbewohner Henning Blunck und Benjamin Dally danke ich dafür, dass sie sich meine Gedanken, meine freudigen Erlebnisse und meinen aus meiner Arbeit resultierend Frust angehört haben, ohne wirklich zu verstehen, was ich eigentlich erforsche.

Schießlich möchte ich auch meinen Eltern und meiner Schwester Laura für das Korrekturlesen und das andauernden Mitfiebern mit dem Fortschritt meiner Arbeit danken. Auch durch sie habe ich während meines Referendariats den nötigen Willen gefunden, meine Arbeit zu einem Ende zu führen, sodass ich jetzt große Freude und Erleichterung verspüre. Natürlich bin ich mir unabhängig davon auch sicher, die besten Eltern und die beste kleine Schwester der Welt zu haben.

Dankbar bin ich auch der Friedrich-Ebert-Stiftung, die mir meine Arbeit nicht nur durch ein Stipendium ermöglicht hat, sondern mich auch mit interessanten anderen Menschen in Kontakt kommen ließ. Die Treffen abseits des eigenen Forschungsbereichs haben mein Leben enorm bereichert.

„Vom vielen Wiegen wird das Schwein nicht fetter.“

1 Einleitung

Das obige Sprichwort ist ein häufig gebrauchtes in der Diskussion über die Sinnhaftigkeit von neu implementierten Formen der standardisierten Schulleistungsmessung in den letzten gut zehn Jahren. Kritiker begegnen damit dem in Folge des „TIMSS-Schocks“ (Weinert, 2002) begonnenen Wandel innerhalb des deutschen Bildungssystems von einer Input- zu einer Outputsteuerung. Diese als „Neue Steuerung“ bezeichnete Steuerung von schulischen Bildungsprozessen erforderte neue Instrumente, um Schulen extern evaluieren zu können. Während sich der gesamtgesellschaftliche Diskurs vorwiegend auf die internationalen Schulleistungsstudien bezieht, richtet sich die Kritik innerhalb der Schulen auch regelmäßig auf die eingeführten nationalen Evaluationsinstrumente. Neben der auf die Einzelschule zielenden Schulinspektion sind dies als Vergleichsarbeiten konzipierte zentrale Lernstandserhebungen (LSE), mit denen Unterrichtsprozesse evaluieren werden sollen.

Der angenommene Wirkungsmechanismus zentraler Lernstandserhebungen

Zentrale Lernstandserhebungen sollen den Unterrichtserfolg in den Fächern Deutsch und Mathematik (in der Grundschule) bzw. Deutsch, Mathematik und optional Englisch oder Französisch (an den weiterführenden Regelschulen) messen. Das dadurch gewonnene Produktwissen soll Lehrkräfte und Fachgruppen anregen und unterstützen, ihren Unterrichtserfolg zu analysieren und den Unterricht ggf. zu modifizieren. Dazu bieten die Projekte VERA3 und Lenstand8/VERA8 Auswertungen an, die objektive Vergleiche zwischen Parallelklassen einer Schule und mit landesweiten Vergleichsgruppen ermöglichen. Aus Sicht der Schulentwicklungsforschung lässt sich der Evaluationsprozess als Ideal mit dem Rahmenmodell der pädagogischen Nutzung von Vergleichsarbeiten von Helmke und Hosenfeld (2005) darstellen. Dieses bildet die Schritte Evaluation – Rezeption – Reflexion – Aktion ab und war Orientierungsrahmen für verschiedene Studien zum Umgang mit Ergebnissen aus zentralen Lernstandserhebungen z.B. (Groß Ophoff, 2013; Hartung-Beck, 2009; Hosenfeld, 2010; Koch, 2011; Müller, 2010; Schneewind, 2007a). Das Rahmenmodell zur pädagogischen Nutzung von Vergleichsarbeiten versteht die aus zentralen Lernstandserhebungen gewonnenen Daten als angebotenes Produktwissen, auf das Lehrkräfte zurückgreifen können, und gelangt nur in diesem Fall zu einer angemessenen Beschreibung der angenommenen Prozesse. Dementsprechend konzentrieren sich bisherige Arbeiten auch auf die Fragen, in welcher Form auf die angebotenen Daten zurückgegriffen wird und wie die Nutzung der Daten unterstützt werden kann.

Gleichzeitig sind LSE nicht nur ein zur Qualitätsentwicklung konzipiertes Instrument, sondern sie werden auch als Instrument zur Rechenschaftslegung gesehen (Peek, 2006). Der administrativen Ebene, aber auch den Erziehungsberechtigten soll durch zentrale Lernstandserhebungen die Möglichkeit gegeben werden, Kontrolle über das Unterrichtsgeschehen auszuüben. Hier liegt ein Input-Prozess-Output-Modell zugrunde, nachdem erstens vom Output – den Testergebnissen zentraler Lernstandserhebungen – auf die Effektivität des Prozesses geschlossen werden kann (Berkemeyer, 2010) und zweitens eine höhere Rechenschaftslegung zu besseren Prozessen führt (Maier, 2010a).

Mögliche Folgen von Rechenschaftslegung

Erfahrungen aus dem anglo-amerikanischen Bereich mit so genannten High-Stake-Tests zeigen, dass eine Zunahme des von Lehrkräften empfundenen Drucks u.a. zu einer speziellen Vorbereitung ihrer Schülerinnern und Schüler auf diese Tests führt (Lind, 2009). Dabei muss diese als Testcoaching bezeichnete Vorbereitung nicht einmal negativ sein. Sollen die den zentralen Lernstandserhebungen zugedachten Funktionen erfüllt sein, dürfen die durch sie gewonnenen Ergebnisse weder manipuliert sein noch dürfen Adressaten des gewonnenen Produktwissens eine Manipulation annehmen. Sinnvoll gestaltetes Testcoaching (im Sinne eines „Familiarity Approach“ – s.u. vgl. Allalouf & Ben-Shakhar, 1998) kann dies sogar unterstützen, indem beispielsweise Testangst abgebaut oder das Selbstvertrauen gesteigert werden. Auch sinnvolle Übungs- und Wiederholungsphasen im Rahmen der Vorbereitung können hilfreich sein, gelten diese im Sinne eines spiralförmigen Lernens doch unabhängig von zentralen Lernstandserhebungen als lernförderlich. Ein zu intensives Testcoaching verfälschte die Messwerte hingegen und verhinderte damit faire Vergleiche, die zentraler Bestandteil der LSE sind. Die zentralen Lernstandserhebungen bilden dann möglicherweise nicht mehr wie vorgesehen den langfristigen Unterrichtserfolg ab, sondern messen lediglich den kurzfristigen Erfolg der Vorbereitungsphase. Darüber hinaus würde bei nur kurzfristigem Lernerfolg wertvolle Unterrichtszeit verschwendet.

Während die Wirkungen von Testcoaching im Zusammenhang mit High-Stake-Tests im angloamerikanischen Raum breit erforscht ist (Abrams & Madaus, 2003; Amrein & Berliner, 2002; Au, 2007; Bond, 1993; Flippo, Becker & Wark, 2000; Kulik, Bangert-Drowns & Kulik, 1984; Messick, 1981; Nichols & Berliner, 2007; Powers, 1998; Ryan, Ryan, Arbuthnot & Samuels, 2007), existieren für den deutschsprachigen Raum nur sehr vereinzelt Studien, die sich mit der Wirkung von Testcoaching auf den Messwert und auf den vorgelagerten Unterricht befassen (Brunner, Artelt, Krauss & Baumert, 2007; Jäger, 2012). Für das Vorbild der LSE in Deutschland, die nationalen Tests in Schweden, muss angenommen werden, dass die Tests dort einmal implementierend bezüglich der Vorgaben, Richtlinien und Lehrpläne wirken, aber auch die Wahl der Unterrichtsinhalte und Arbeitsformen beeinflussen (Erickson & Lander, 2007). Auch in Schweden werden folglich Testcoaching-Effekte befürchtet und erkannt. Zu den seit 2004 in Nordrhein-Westfalen existierenden LSE existieren bisher keine Studien, die tatsächlich empirisch Daten erhoben haben. Weder die Ergebnisse der wenigen

deutschen Studien noch die Ergebnisse der vielen Studien aus dem anglo-amerikanischen Raum lassen sich ohne weiteres auf die LSE übertragen. Für die Ergebnisse der U.S.-amerikanischen Forschung liegt dies u.a. am anderen kulturellen Rahmen. Gravierender ist aber, dass eine Einordnung als High-Stake-Test oder Low-Stake-Test nicht vorgenommen werden kann: Erstens sind die Ergebnisse der Tests im Gegensatz zu SAT[1] und GRE[2] nicht nur zusätzlich auch für die ausbildenden Institutionen von Bedeutung, sondern die Ergebnisse der LSE sind es in erster Linie.[3]

Zweitens ist darüber hinaus zu vermuten, dass viele Lehrerinnen und Lehrer sich der Folgen des Testcoachings für die Möglichkeit, die LSE zur Qualitätsentwicklung zu nutzen, nicht bewusst sind, oder dass verfälschte Testergebnisse ganz im Interesse einiger Lehrerinnen und Lehrer liegen, um die Praxis der Rechenschaftslegung zu verhindern. Gerade die Sicht auf die LSE als ein Instrument der Rechenschaftslegung scheint besonders die schädlichen Varianten des Testcoachings – auch „Teaching for the Test“ genannt – zu begünstigen (van Ackeren & Bellenberg, 2004; Bonsen & von der Gathen, 2004). Aber auch das fehlende Verständnis für den Wirkungsmechanismus zentraler Lernstandserhebungen ohne sabotierendes Motiv scheint bei vielen Lehrkräften den Gedanken manifestiert zu haben, eine Vorbereitung auf LSE gehöre zwingend dazu (Hahn, 2008). Eine vorab vom Autor dieser Arbeit durchgeführte Interviewstudie im Jahr 2008 lässt für Lehrkräfte aus Nordrhein-Westfalen vermuten, dass der vorgelagerte Unterricht vor der VERA8-Erhebung jedes Jahr durch umfangreiche Vorbereitungsphasen geprägt ist. Entgegen des obigen Sprichworts scheint das zusätzliche Wiegen durch zentrale Lernstandserhebungen durchaus dazu zu führen, dass Schülerinnen und Schüler zumindest mehr – um im Bild zu bleiben - gefüttert werden. Es bleibt aber offen, ob dabei sinnvolles Vorbereitungsverhalten an den Tag gelegt wird.

Lehrkräfte als Experten für das Unterrichten und die Innovation von Unterricht

Zehn Jahre vor dem durch internationale Schulleistungstests ausgelösten Paradigmenwechsel von der Input- zur Outputsteuerung vollzog sich in der Lehrer-Forschung bereits der Paradigmenwechsel vom Prozess-Produkt-Paradigma zur individualpsychologischen Sichtweise des Lehrer-Expertenansatzes. Allgemein sind Lehrkräfte demnach autonom Handelnde, die auf Wissen und Können aus dem Studium und praktischer Erfahrung zurückgreifen, um die an sie gestellten Aufgaben zu bewältigen (Bromme, 2008; Dann, 2008). Dieser enge Kompetenzbegriff wurde mittlerweile durch einen Kompetenzbegriff im weiteren Sinne ersetzt, sodass die Modelle auch Überzeugungen (als Werthaltungen, subjektive Theorien und Ziele) sowie motivationale Orientierungen und

[1] Scholastic Achievement Test (heute: SAT Reasoning Test).
[2] Graduate Record Examination.
[3] Dies zeigt im Übrigen, wie wenig die eindimensionale Klassifikation high-stake und low-stake ausreicht, um angemessen den jeweiligen Stellenwert von Tests zu beschreiben. Der Stellenwert, den die LSE für die Schülerinnen und Schüler einer Klasse oder Schule besitzen, muss nicht mit dem Stellenwert identisch sein, den die Lehrkräfte der Schule oder Klasse den LSE zuschreiben.

selbstregulative Fähigkeiten (das Berufserleben und Kompetenz- und Kontrollüberzeugungen) berücksichtigen (Baumert & Kunter, 2006). In dem Bereich durchgeführte Studien zeigen einen naheliegenden Einfluss von Wissen und Können auf die Unterrichtsqualität (Baumert et al., 2010; Blömeke, Kaiser, Döhrmann & Lehmann, 2010; Blömeke & König, 2010), aber auch ein Zusammenhang der motivationale Orientierungen und selbstregulative Fähigkeiten zur Unterrichtsqualität konnte nachgewiesen werden (Klusmann, 2008).

Der Umgang mit zentralen Lernstandserhebungen stellt eine derjenigen Aufgaben dar, mit denen Lehrkräfte konfrontiert werden und denen sie im Rahmen ihrer individuellen Möglichkeiten professionell begegnen. Als Aufgabe kann aber dabei nicht nur die Nutzung des angebotenen Produktwissens verstanden werden. Ebenso erscheint eine mögliche Vorbereitung auf zentrale Lernstandserhebungen manchen Lehrkräften als Aufgabe, die es innerhalb der Unterrichtsumgebung zu bewältigen gilt. Somit ist ein Zusammenhang zwischen der jeweiligen Lehrer-Handlungskompetenz und dem durchgeführten Testcoaching anzunehmen.

Konzeption der Arbeit

Die vorliegende Arbeit nimmt die Sicht des Lehrer-Expertenansatzes ein, um das Vorbereitungsverhalten der Lehrkräfte zu untersuchen. Dazu wurden 2010 parallel zwei Fragebogen-Studien unter den Gymnasiallehrkräften in Nordrhein-Westfalen durchgeführt, die zum Durchführungszeitpunkt von Lernstand8/VERA8 in einer achten Klasse unterrichteten. Beide Studien erhoben eine möglicherweise durchgeführte Vorbereitung auf den entsprechenden Test. Es wurden Daten gesammelt zur aufgewendeten Unterrichtszeit, den durchgeführten Maßnahmen, den Inhalten von möglichen Übungs- und Wiederholungsphasen, der Vermittlung von Test-Wiseness-Strategien und dem Einsatz von unterstützenden Lernmaterialien. Außerdem wurden u.a. mögliche Veränderungen – auch mit Blick auf Lernstand8/VERA8 – des Unterrichts und im Vorbereitungsverhalten im Vergleich zu früheren Jahren erfragt und auch die außerschulische Vorbereitung erhoben. Dadurch sollte es ermöglicht werden, zu einem umfassenden Bild des durchgeführten Testcoachings im Vorfeld von Lernstand8/VERA8 zu gelangen.

Um das so gewonnene Bild der Testcoachingqualität mit der Lehrer-Professionalität zu verknüpfen, wurden in beiden Studien darüber hinaus entsprechende Daten gesammelt und zwei differentielle Ansätze betrachtet. Beide Ansätze nutzen die Methode der latenten Klassenanalyse. In Studie A lag der Schwerpunkt darauf, die Qualität der Vorbereitung über Überzeugungen i.S. von Werthaltungen, subjektive Theorien und Ziele sowie motivationale Orientierungen und selbstregulative Fähigkeiten zu erklären. Der erste differentielle Ansatz nutzt folglich einen Ausschnitt aus dem Modell der Lehrer-Handlungskompetenz nach Baumert und Kunter (2006). Dieses wird im Theorieteil diskutiert und im Hinblick auf die pädagogische Nutzung von Produktwissen aus Evaluationen erweitert. Die Vorbereitung auf

Lernstand8/VERA8 ist aus dieser Perspektive vorwiegend ein Teil des normalen Unterrichts. Entsprechend besteht hier die Grundannahme, dass sich Qualitätsunterschiede auch analog zu anderem Unterricht erklären lassen.

Studie B hingegen fokussierte auf die Elemente der pädagogischen Nutzung von Ergebnissen aus Vergleichsarbeiten. Die Vorbereitung wird i.S. des Rahmenmodells von Helmke und Hosenfeld als Pro-Aktion behandelt und als Re-Aktion auf ein angekündigtes Feedback aufgefasst. Vor dem Hintergrund der Feedback-Interventions-Theorie (Kluger & DeNisi, 1996) sind hier Bedingungen erhoben worden, die den Grad beeinflussen, wie weit man bereit ist, sich mit angebotenem Feedback auseinanderzusetzen. Außerdem wurde der bisherige Umgang mit dem angebotenen Produktwissen per Selbstauskunft gemessen. Dadurch konnten Nutzungstypen bestimmt werden, die hier Qualitätsunterschiede der Vorbereitung erklären sollten.

Entsprechen den beiden differentiellen Ansätzen und dem Begriffspaar Qualitätsentwicklung-Rechenschaftslegung weist diese Arbeit einerseits die klassische Struktur von theoretischem Teil und empirischem Teil auf, besitzt aber innerhalb der einzelnen Kapitel immer wieder parallele Stränge.

Kapitel 2 befasst sich mit standardisierter Schulleistungsmessung. Dargestellt werden hier die verschiedenen Formen standardisierter Schulleistungsmessung, zu denen auch zentrale Lernstandserhebungen und im Speziellen zentrale Vergleichsarbeiten gehören. Dargestellt werden erstens die Intentionen von standardisierter Schulleistungsmessung aus steuerungstheoretischer Sicht, die Wirkungen von standardisierter Schulleistungsmessung als Rechenschaftslegung und die datengestützte Qualitätsentwicklung durch standardisierte Schulleistungsmessung. Zweitens wird eine Darstellung zentraler Lernstandserhebungen in Deutschland gegeben und es werden die Leistungsmessung im Bundesgebiet, die Anlage von VERA8 in der Bundesrepublik allgemein und die spezielle Umsetzung in Nordrhein-Westfalen sowie die Steuerungslogik und Wahrnehmung zentraler Vergleichsarbeiten in Deutschland betrachtet.

Im Kapitel 3 geht es um Testcoaching. Zuerst werden die testtheoretischen Grundlagen skizziert, wobei dieses wirklich nur eine Skizze ist und eine umfassende Betrachtung weitere Modelle als nur das dort beispielhaft betrachtete lineare Modell berücksichtig werden müsste. Es wird weiter ein Vorschlag für eine sinnvolle Definition von “Testcoaching“ gegeben und es werden drei verschiedene Herangehensweisen bei der Testvorbereitung differenziert. Anschließend folgt eine Übersicht über den Forschungsstand zur Wirkung von Testcoaching auf die Testergebnisse. Die zugrunde gelegten Studien beziehen sich alle auf Testcoaching im Rahmen von PISA[4], SAT und GRE. Zum Schluss des Kapitels werden die theoretischen Möglichkeiten diskutiert, auf Lernstand8/VERA8 in Nordrhein-Westfalen

[4] Programme for International Student Assessment – eine alle drei Jahre von der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) durchgeführte internationale Schulleistungsstudie

vorzubereiten und wesentliche Ergebnisse der 2008 durchgeführten Interviewstudie skizziert.

Kapitel 4 schlägt dann den Bogen zum Lehrer-Expertenansatz, der zuerst ausführlich beschrieben wird. Es wird weiter das Modell der Lehrer-Handlungskompetenz nach Baumert und Kunter im Zusammenhang mit weiteren Forschungsergebnissen diskutiert. Mit dem Abschnitt über die Lehrerpersönlichkeit als handlungsbestimmendes Element bei Rechenschaftslegung und Qualitätsentwicklung wird auch der andere Strang weiter verfolgt. Aus beiden Strängen zusammen wird dann ein erweitertes Modell der Lehrer-Handlungskompetenz hergeleitet. Diese drei Kapitel bilden zusammen den Theorieteil.

Der empirische Teil verbindet die im Theorieteil dargelegten vorherigen Befunde, indem zuerst Forschungsfragen und Hypothesen formuliert werden. Dies geschieht in Kapitel 5. Das Kapitel 6 ist das Methodenkapitel. Hierin sind das Design der Studie beschrieben, die Untersuchungsinstrumente begründet und die Fragebogen-Items ausführlich vorgestellt. Auch die Datenauswertung wird reflektiert. In Kapitel 7 sind ausführlich alle Befunde aus beiden Studien dokumentiert. Dabei werden zuerst nur die Befunde zum Testcoaching dargestellt. Anschließend wird beschrieben, wie letztendlich die beiden Modelle für die beiden differentiellen Ansätze gewonnen wurden. Im dritten Abschnitt des Kapitels werden dann die Befunde zum Testcoaching im Zusammenhang mit dem ersten Modell und im vierten Abschnitt die Befunde im Zusammenhang mit dem zweiten Modell berichtet.

Diese Arbeit schließt mit Kapitel 8. Dieses bietet zuerst eine Diskussion der Ergebnisse vor dem Hintergrund der Forschungsfragen und Hypothesen. Auch hierbei existiert wiederum die Aufteilung in parallele Stränge. Danach werden im Nachhinein sichtbar gewordene Einschränkungen der Gesamtergebnisse dieser Arbeit diskutiert. Den endgültigen Abschluss bilden ein Gesamtfazit und mögliche Perspektive für Praxis und Forschung. Damit wird auch mittels des Begriffspaars Qualitätsentwicklung/Rechenschaftslegung noch einmal ein Rückgriff auf die in Kapitel 2 skizzierten Forschungszweige vorgenommen, die sich mit der Steuerungslogik von zentralen Lernstandserhebungen befassen.

Theorieteil

2 Standardisierte Schulleistungsmessung

2.1 Standardisierte Schulleistungsmessung in Europa, Neuseeland und den USA: zwischen Rechenschaftslegung und Qualitätsentwicklung

Der erste Teil dieses Kapitels beschäftigt sich mit standardisierter Schulleistungsmessung in Europa, Neuseeland und den USA. Auch zentrale Lernstandserhebungen[5] im Sinne von VERA3 und VERA8 fallen unter die Kategorie der standardisierten Schulleistungsmessung. Da vergleichende Lernstandserhebungen im deutschsprachigen Raum erst vor wenigen Jahren eingeführt wurden, soll der Blick über den deutschen Tellerrand hinaus helfen, die Intentionen und Steuerungsideen hinter zentralen Lernstandserhebungen zu verstehen. Die Unterschiede der Instrumente in Deutschland zu den in anderen Staaten bereits längere Zeit umgesetzten ähnlichen Instrumenten, aber auch die Effekte sollen auf diese Weise deutlich werden. Zwei besondere Schwerpunkte liegen bei dieser Darstellung auf den Funktionsweisen der Rechenschaftslegung (als „Pressure"-Ansatz charakterisierbar – van Ackeren, 2003) und der Qualitätsentwicklung durch Evaluationsunterstützung („Support-Ansatz" - ebenda). Die Formel von „Pressure and Support" geht ursprünglich auf Huberman und Miles (1984) zurück, die in „richtungsweisenden Untersuchungen" (Kempfert & Rolff, 2005, S. 19) Bedingungen von Implementationen herausgearbeitete haben und u.a. feststellten, dass Implementationen wahrscheinlicher gelingen, wenn es zu einem Zusammenspiel von Druck durch Rechenschaftslegung und Unterstützungsangeboten bei der Qualitätsentwicklung kommt.[6] Die strikte Trennung in „Pressure"- und „Support"-Konzeptionen ist schon früher von van Ackeren (2003) zugunsten einer dreiwertigen Einteilung mit Mischkategorie aufgegeben worden, da die in verschiedenen Staaten realisierten Umsetzungen in keine eindeutige Klassifikation dieser Art einsortiert werden können. Durch in den letzten zehn Jahren vorgenommene Veränderungen gilt dies auch für vorher eindeutig klassifizierbare Ansätze. Wenn Staaten in die ursprünglichen Gruppen

[5] Im deutschen Sprachraum werden die Begriffe „Kompetenztest", „zentrale Lernstandserhebung", „Orientierungsarbeit" und „zentrale Vergleichsarbeiten" häufig synonym verwendet. In dieser Arbeit soll „Lernstandserhebung" als Oberbegriff für alle standardisierten Verfahren der Schulleistungsmessung dienen, die nicht am Ende eines formalen Ausbildungsabschnittes durchgeführt werde. „Vergleichsarbeiten" soll speziell diejenigen Lernstandserhebungen bezeichnen, bei denen der Vergleich mit anderen Lerngruppen und Lernenden eine entscheidende Rolle spielt. Die Bezeichnungen „Orientierungsarbeiten" (Baden-Württemberg) und „Kompetenztests" (Thüringen) werden dabei wiederum nur als Name dieser standardisierten Schulleistungsmessung in bestimmten Bundesländern verwendet.

[6] Die Übersetzung der Formel als „Druck und Zug" ist nicht nur unpräzise, weil „Unterstützung" nicht genannt wird, wie Kempfert und Rolff (2005) schreiben, in der Übersetzung drücken sich auch unterschiedliche Konzeptionen aus, denn „Zug" lässt Schulen in ihrer passiven Rolle, während „Support" Unterstützungs-*Angebote* meinen kann.

einsortiert werden, ist dies also symbolisch, nicht beispielhaft. Die globale Betrachtung soll vielmehr eine Grundlage für die in den weiteren Teilen folgende Analyse des Instruments VERA8 bilden.

2.1.1 Intentionen von standardisierter Schulleistungsmessung aus steuerungstheoretischer Sicht

Als standarisierter Schulleistungsmessung (SSL) sollen im engeren Sinne Tests klassifiziert werden, die standardisierte, zentrale und auf nationaler Ebene [7] durchgeführte Leistungsmessungen darstellen.[8] Sie können der (summativen oder formativen) Beurteilung sowie Qualitätsentwicklung dienen (Parveva, Coster & Noorani, 2009). Derart verstandene Schulleistungsmessung umfasst sowohl zentrale Lernstandserhebungen als auch zentrale Abschlussprüfungen und grenzt sich von anderen standardisierten Formen ab wie Large Scale Assessments, die als ebenfalls standardisierte Messverfahren auch wichtige Beiträge für das Systemmonitoring liefern können (vgl. auch Kühle & van Ackeren, 2012). Ein wesentliches Charakteristikum ist neben dem einmaligen Messzeitpunkt die enge Bindung an zielorientierte Vorgaben wie Standards oder Lehrpläne (auch: „Curriculum Alignment“) statt an kurze Lernphasen, wodurch sich zentrale Lernstandserhebungen und zentrale Abschlussprüfungen von den Verfahren der fortlaufenden Leistungsmessung (z.B. so genannte „schriftliche Arbeiten“) unterscheiden. Nach Bonsen, Büchter und Peek (2006) zeichnet sich die SSL außerdem dadurch aus, dass sie gültige und aussagekräftige Ergebnisse bietet und Vergleiche der Lernergebnisse zwischen Lerngruppen zulässt. SSL kann verpflichtend, aber auch fakultativ sein. Sie ist als jahrgangstufenweise Vollerhebung oder auch stichprobenartig konzipiert.

Die standarisierte Schulleistungsmessung kann nach Parveva, De Coster und Noorani (2009) in drei Kategorien eingeteilt werden, wobei es in der konkreten Umsetzung aber häufig zu Mischformen kommt. Die SSL kann im Sinne einer *summativen Beurteilung* genutzt werden, um den Lernstand von Schülerinnen und Schülern am Ende eines Schuljahres oder am Ende eines Schulabschnittes zu erheben, und besitzt in dem Fall eine erhebliche Auswirkung auf die Bildungslaufbahn der Schülerinnen und Schüler. Die Prüfungen haben zusätzlich häufig eine zertifizierende Funktion und werden daher bis auf wenige Ausnahmen nur am Ende der Pflichtschulbildung durchgeführt. Beispiele hierfür sind in Deutschland das Zentralabitur oder die Zentralen Abschlussprüfungen am Ende der Vollzeitschulpflicht. Ebenfalls auf einzelne Schüler und Schülerinnen bezieht sich eine *formative Nutzung* von SSL. Die Ergebnisse der Tests dienen in diesem Fall der Diagnose von Lernständen, um den

[7] „Nationale Ebene“ meint auch zentrale Lernstanderhebungen, bei denen die konkrete Ausgestaltung in der Verantwortung von Bundesländern oder Provinzen liegt.

[8] Test werden auch dann als standardisiert bezeichnet, wenn sie zentral gestellte Prüfungsaufgaben beinhalten und nach zentralen Vorgaben korrigiert werden. Es wird an dieser Stelle beispielsweise nicht explizit gefordert, dass alle Aufgaben zentral gestellt werden oder „Blind-Korrekturen“ vorgenommen werden.

individuellen Lernprozess steuern zu können, indem Lernangebote dem erhobenen Lernstand angepasst werden können. SSL dieser beiden Kategorien setzt einen direkten Bezug zwischen der Nutzung der Testergebnisse und den getesteten Schülerinnen und Schülern voraus. Die dritte Gruppe von Tests als standarisierte Schulleistungsmessung legt den Fokus nicht auf einzelne Schüler und Schülerinnen, sondern liefert im idealen Fall Daten zum Systemmonitoring (auf der Makro- und Mesoebene). Zentrale Vergleichsarbeiten wie VERA3 oder VERA8 sollen (auch) diese Funktion ermöglichen. Sie werden entsprechend ähnlich zu Large-Scale-Assessments derart angelegt, dass sie ebenfalls Rasch-skalierbar sind und theoretisch klare Aussagen über den Erfolg von Bildungspolitik zulassen (Maier, Metz, Bohl, Kleinknecht & Schymala, 2012). Alternativ können auch Daten auf der Mikro- bzw. Klassenebene gewonnen werden. Die Daten lassen sich dann zur externen Evaluation von Schule und Unterricht nutzen. Auf der Klassenebene wird mit dieser Form der SSL die Unterrichtsqualität bzw. ein Teil der Leistung der Lehrkraft gemessen (Parveva et al., 2009). Um externe Evaluation handelt es sich, da die Messinstrumente zentral und damit extern konzipiert (und in der Regel auch initiiert) werden. Auch hier können zentrale Vergleichsarbeiten wie VERA3 oder VERA8 als Beispiel dienen (Maier et al., 2012). Folglich soll die SSL gleichzeitig eine *Unterstützung der Qualitätsentwicklung und Rechenschaftslegung* ermöglichen.[9]

Standardisierter Schulleistungsmessung wird bereits für die letzte dieser drei Gruppen mit verschiedensten Zielen[10] verbunden. Zur gleichen Zeit sollen mit SSL je nach Konzeption Aufgaben der Bildungsadministration umgesetzt werden und Lehrkräften und Schulen eine Weiterentwicklung ermöglicht werden. Maritzen (2008) spricht davon, mit den Daten aus externer Evaluation[11]

- Qualitätssicherung zu betreiben,
- wissenschaftliche Erkenntnis über Wirksamkeit schulischer Arbeit zu erlangen,
- Impulse für die Schulentwicklung zu geben,
- Lehrkräften und Schulen ein Feedback zu geben/zu spiegeln und
- Unterstützung einzelner Schulen zu ermöglichen.

Kühn nennt (auch für die ersten beiden Formen gegenüber nicht-standardisierten Schulleistungsmessungen) außerdem die Entlastungsfunktion als Ziel. Lehrkräfte müssen Tests der standardisierten Schulleistungsmessung in manchen Staaten zwar weiterhin selbst (oder kreuzweise) auswerten, sie werden aber

- von der Konzeption von Tests entlastet (Kühn, 2010).

[9] Zur Unterscheidung zwischen verschiedenen Formen s. nächste Abschnitte.

[10] Die Ziele werden stellenweise als „Funktionen“ bezeichnet. Zur Unterscheidung von den aus lerntheoretischer Sicht hergeleiteten Funktionen wird hier von „Zielen“ gesprochen.

[11] Diese Sammlung von Intentionen betrachtet die Instrumente mit dem Fokus auf Schulinspektionen aus der Steuerungsperspektive Maritzen (2008), ist aber auf die Verfahren standardisierter Schulleistungsmessung zu übertragen.

Weiterhin erhofft man sich auch Vorteile für die direkten (Erziehungsberechtigte und Schüler) und indirekten (Bürger insgesamt) Abnehmer des Bildungssystems. SSL sollen daher auch

- über die Leistungen der Einzelschulen informieren und ggf. Wettbewerb ermöglichen (Tresch, 2007).[12]

Neben dieser steuerungstheoretischen Sicht auf die Intentionen von SSL können ihnen auch aus lerntheoretischer Sicht eine Reihe von Zielen zugesprochen werden. Je nach konkreter Konzeption sollen sie dann

- die im Unterricht bereits erworbenen Kompetenzen der Schülerinnen und Schüler darstellen (deskriptive Funktion), um den Kompetenzstand mit den Zielvorgaben (Standards) vergleichen zu können,
- die Intentionen der Lehrpläne und Standards illustrieren (implementierende Funktion), indem die Testaufgaben (im Sinnes eines Feed-forward, aber auch zur Unterstützung der Lehrkräfte) darstellen, welche Kompetenzen erworben werden müssen,
- Anhaltspunkte zur Diagnose sowohl für einzelne Schülerinnen und Schüler als auch für Klassen und Jahrgangsstufen bieten (diagnostische Funktion), um eine Adaption des Unterrichts vornehmen zu können,
- Signalcharakter zur Gestaltung von Unterricht und Leistungsbewertung haben (erste innovierende Funktion[13]), indem die Testaufgaben selbst von Leistungs- und Lernaufgaben durch die Lehrkräfte transformiert und in den Unterricht integriert werden und indem das kriteriale Auswertungsprinzip veranschaulicht wird,
- Unterstützung bei der internen Evaluation[14] anbieten (zweite innovierende Funktion), indem standarisierte Instrumente zur Leistungsmessung und Leistungsdaten zur Verfügung gestellt werden [und]
- einen Beitrag zur Entwicklung empirisch abgesicherter Kompetenzniveaus liefern, die zur Weiterentwicklung der Lehrpläne und Standards genutzt werden können (entwickelnde Funktion) (Heymann & Pallack, 2007; Parveva et al., 2009 – Durchzählung vom Verfasser).

In den meisten Fällen werden die SSL auch zur Gewinnung von Individualdaten genutzt, d.h. die Leistungen, die Schülerinnen und Schüler bei SSL gezeigt haben, sollen

[12] Siehe auch weiter unten.

[13] Hier drückt sich wie auch in den zweiten innovierenden Funktionen die enge Beziehung zwischen der steuerungstheoretischen und der lerntheoretischen Perspektive aus. Beide Perspektiven drücken den Wunsch aus, das Lernen durch standardisierte Schulleistungsmessung in der Schule zu verbessern. Während die steuerungstheoretische Perspektive allerdings diskutiert, wie SSL als ein Element von Steuerung wirkt und eingebettet ist, ist der Blick aus lerntheoretischer Sicht vor allem auf die verschiedenen Stellschrauben des Lernens fokussiert.

[14] Zum Unterschied zwischen interner und Selbstevaluation siehe Berkemeyer und Müller (2010).

- zusätzlich als Kriterium zur Schullaufbahnempfehlungen genutzt werden können (prognostische Funktion).

Diese Individualdiagnostik kann sowohl mit fakultativen Tests (z. B. in den Niederlanden) als auch mit obligatorischen Tests (z.B. im Vereinigten Königreich von Großbritannien) geschehen (Burkard & Peek, 2004; Helmke & Hosenfeld, 2003; Parveva et al., 2009).

SSL als Mittel zur Realisierung verschiedenster Wünsche wird aber auch wegen dieser Ansammlung an Intentionen kritisiert (beispielshaft O`Day, 2004; Parveva et al., 2009). Kühle und Peek warnen, dass diese Fülle an Zielen im Sinne einer „Funktionsüberfrachtung“ mit dazu beiträgt, diese Ziele schlussendlich nicht zu erreichen (Kühle & Peek, 2007). Wird SSL zusätzlich oder vor allem zur summativen Beurteilung benutzt, kommen außerdem noch die Allokations- und Selektionsfunktion, die Anreiz- und Disziplinierungsfunktion und schließlich die Sozialisationsfunktion hinzu, die auch den traditionellen Formen der schulischen Leistungsmessung zugesprochen werden. Schon für diese Funktionen der gewöhnlichen Leistungsmessung und -bewertung kann angenommen werden, dass es unmöglich ist, allen gerecht zu werden (Sacher, 2009). Auch wird in Frage gestellt, ob selbst standardisierte Instrumente in der Lage sind, die Erwartungen einer Prognose für den weiteren schulischen Weg zu erfüllen (Bunting & Mooney, 2001).

Standardisierte Schulleistungsmessung ist zu einem wichtigen Instrument für politische Steuerung im Bildungswesen geworden. Dabei treten Steuerungseffekte auch dann auf, wenn diese nicht die Hauptintention der eingeführten SSL sein sollen. Wenngleich zentrale Prüfungen beispielsweise in den Vereinigten Staaten von Amerika und in einigen europäischen Staaten (wie Island oder Großbritannien) eine Jahrzehnte dauernde Tradition aufweisen können, hat SSL insbesondere in den letzten zwanzig Jahren flächendeckend in den europäischen Staaten als Instrument zur Beurteilung und Qualitätsentwicklung an Bedeutung gewonnen. Dabei ist gleichzeitig eine Tendenz von zentralen Prüfungen zur Beurteilung von Schülerinnen und Schülern hin zu zentralen Lernstanderhebungen als Instrument zur Qualitätsentwicklung zu beobachten (Mons, 2009; Parveva et al., 2009).

Die flächendeckende Einführung von standardisierter Schulleistungsmessung in Europa führt man nach Maag Merki (2010) auf drei Gründe zurück:

(1) Seit den neunziger Jahren herrscht in der Schulentwicklungsforschung die Ansicht, Schulentwicklung bedürfe eines Mischsystems aus Selbst- und Fremdsteuerung. Die selbstständige Einzelschule als lern- und handlungsfähige Institution müsse dabei im Mittelpunkt stehen, da „von oben“ (von der administrativen Ebene) verordnete Innovationen in den Schulen nicht ausreichend umgesetzt wurden (Rolff, 2006). Innerhalb der letzten zwanzig Jahre wechselte zusätzlich der Fokus von verwaltungstechnischen Reformen, die den Schulen mehr Autonomie gaben, hin zu der Vorstellung, das Unterrichtsgeschehen müsse seine Zieldefinition an operationalisierten, zentral

vorgegebenen Standards ausrichten (*standardorientierte Steuerung*[15]). Dabei wird die häufig *output-orientierte Steuerung* durch andere Reformvorhaben begleitet. Beispielsweise werden mit den in einigen Staaten eingeführten Standards neue inhaltliche Vorgaben gemacht oder die im Rahmen der Rechenschaftslegung implementierten zentralen Lernstandserhebungen werden durch didaktische Erläuterungen der Testaufgaben ergänzt (Maag Merki, 2010; Maier, 2009).

(2) Diese schulische Qualitätssicherung und -entwicklung rückte nach Bildungsschocks, die beispielsweise durch TIMSS 1997 und PISA 2000 in Deutschland und Österreich ausgelöst wurde, ins Bewusstsein der politischen Akteure (Peek, Pallack, Dobbelstein, Fleischer & Leutner, 2006). Voraussetzung für die durch internationale Schulvergleichsstudien ausgelösten Reformen ist nach Mons (2009) und Maier (2009) der Bewusstseinswandel hin zu quantitativ messbaren Lerninhalten. Altrichter (2010) spricht von einem Prinzip der Evidenzbasierung in Bildungspolitik und Schulentwicklung, welche Entscheidungen aufgrund von geprüften Informationen fällt und ihrer Umsetzung empirisch evaluiert (*evidenzbasierte Steuerung*). Dabei kommt es gleichzeitig zu einer fachlichen Unterrichtsfokussierung (Altrichter, 2010), die zu Ungunsten von schulischen Inhalten des sozialen Miteinanders führe, da hier Veränderungen einfacher messbar sind. Unterstützend wirkt hier die seit den achtziger Jahren zu beobachtende Vorstellung von Fertigkeiten und Fähigkeiten als humanes Kapital im Sinne einer ökonomischen Humankapitaltheorie − eine Sichtweise, die als eine Folge der Transformation von Industriegesellschaften gesehen werden kann (Maier, 2009; Mons, 2009).

(3) Es lässt sich in den letzten Jahrzehnten eine Verschiebung der Macht und Kontrolle von staatlichen Akteuren zu den Bürgern als Akteure beobachten. Damit einher geht ein größeres Interesse der bürgerlichen Akteure, insbesondere der Erziehungsberechtigten, an Rechenschaftslegung der Schulen gegenüber der Öffentlichkeit (ebenda). Maier merkt hierzu an, dass testbasierte Schulreformen in der Öffentlichkeit auch wegen ihrer Einfachheit eine große Akzeptanz besitzen. Dies führt möglichweise dazu, dass eine Diskussion dieser Reformidee in der Medienöffentlichkeit als auch im politischen Bereich unterbleibt (Maier, 2009).

Die Steuerungswirkung von zentralen Lernstandserhebungen und zentralen Abschlussprüfungen wird in der Bildungsforschung von drei Theoriezweige untersucht: von der vorwiegend soziologisch geprägten Educational Governance-Perspektive, von der Schulqualitätsforschung und von der Schulentwicklungsforschung (Maag Merki, 2010; Mons, 2009). Der Blick aus den drei Perspektiven soll hier kurz skizziert werden. Alle drei Perspektiven liefern Elemente für die Analyse des Vorbereitungsverhaltens, sodass die

[15] Diemer und Kuper weisen darauf hin, dass „standardorientierte Steuerung“, „output-orientierte Steuerung“ und „evidenzbasierte Steuerung“ *die* drei Elemente der Neuen Steuerung sind, sich inhaltlich aber natürlich unterscheiden Diemer und Kuper (2011). Streng genommen handelt es sich bei standardorientierter Steuerung eher um eine input-orientierte Steuerung Maag Merki (2010), Output-Orientierung und Evidenzbasierung sind inhaltlich aber verwoben und Evidenzbasierung braucht Ziele, die durch die Standards gegeben werden. In dieser Arbeit folgen differenziertere Betrachtungen daher erst an späterer Stelle.

Skizzen im Sinne einer ersten Erkundung eine Berechtigung finden. Für eine detaillierte Auseinandersetzung ist in dieser Arbeit allerdings kein Raum. Dazu muss auf die zitierten Arbeiten verwiesen werden.

Aus der **Educational Governance-Perspektive** lässt sich das Verhältnis von Lehrkraft und Einzelschule zu den Bürgerinnen und Bürgern bzw. den Schülerinnen und Schülern und Erziehungsberechtigte (einfachhalber beide zu Staat zusammengefasst) mit der Prinzipal-Agent-Theorie (als eine New Public Management-Theorie) beschreiben (Clausen, 2007). Die Theorie geht ursprünglich auf Ross (1973) zurück. Mit der Prinzipal-Agent-Theorie werden Vertragsabhängigkeiten beschrieben, in denen der Prinzipal auf die Leistungserbringung des Agenten angewiesen ist, gleichzeitig aber an einem Informationsdefizit leidet. Die Theorie nimmt an, dass beide Parteien an einer Nutzenmaximierung zum eigenen Vorteil interessiert sind. Die Ziele der Parteien sind dabei erst einmal nicht kohärent (Ross, 1973). Lehrkräfte und Einzelschule treten hier als Agenten[16], als Beauftragte des Staates, der Staat als Prinzipal bzw. Beauftragender auf. Der Staat vergibt an den Agenten den Bildungsauftrag. Dabei liegt die Art und Weise der Auftragsausführung innerhalb der Einzelschule bzw. innerhalb des Unterrichts in der Eigenverantwortung des Agenten. Um das dadurch herrschende Informationsdefizit (eigentlich: Kontrolldefizit) zu beheben, werden Kontroll- und Anreizmechanismen implementiert (z.B. Jensen & Meckling, 1976; Kieser & Ebers, 2006). Der Agent wiederum kann das Auftragsverhältnis kündigen, und zwar sichtbar oder auch nur latent (Kussau, 2007). Im Bildungsbereich sind Verfahren wie standardisierte Schulleistungsmessungen oder Schulinspektionen mögliche Kontrollmechanismen. Anreizmechanismen sind selten vorgesehen (z.B. „der Deutscher Schulpreis“). Die mit der New Public Management-Idee angestrebten Ziele (Schaffung von Marktsituationen, Selbstständigkeit der Verwaltungseinheiten, Transparenz der Verwaltung, Effizienzsteigerung und Kontrolle) werden allerdings sehr unterschiedlich verstanden und umgesetzt: Bereits angesprochen wurde die Veröffentlichung von Ergebnissen aus standardisierter Schulleistungsmessung. Demnach fungiert in Deutschland (und den meisten anderen europäischen Staaten) der Staat als Prinzipal, während in den U.S.A. und in Großbritannien tatsächlich die Öffentlichkeit als Rezipient der Ergebnisse auftreten und Erziehungsberechtigte beispielsweise auf Grundlage dieser Informationen eine Schulwahl treffen[17] (Clausen, 2007). Eine Marktsituation wird ebenfalls entsprechend unterschiedlich zugelassen (Maag Merki, 2010). Inwiefern die Ziele Kontrolle und Effizienzsteigerung erreicht werden, hängt stark von den implementierten Instrumenten ab. Wesentlich sind dabei die klare Definition der Ziele und die Validität der Kontrollinstrumente. Anhaltspunkte liefern dazu die zahlreichen Studien zur Wirkung von High-Stake-Tests (s.u.), die aber wiederum nur in einigen Staaten Verwendung finden (s. auch (2.1.2)).

[16] Kussau (2007) meint als „institutionelle Akteure“, die an den politisch definierten Auftrag gebunden sind.

[17] Auch in Deutschland stehen Erziehungsberechtigten Daten aus an den Schulen ihrer Kinder durchgeführter SSL theoretisch zur Verfügung. Die Daten werden aber nicht flächendeckend öffentlich gemacht und von Erziehungsberechtigten nur auf Individual- oder Klasseneben nachgefragt.

Im Sinne der **Schulqualitätsforschung** bieten sich nach Fends Theorie der Schule (Fend, 2008, 2009) verschiedene theoretische Positionen an. Maier hält davon besonders die systemtheoretische Perspektive inklusive des oft beschriebenen Technologiedefizits[18] sowie eine neo-institutionelle Perspektive als Diskussionsfolie für fruchtbar (Maier, 2009).[19] Das Technologiedefizit der systemtheoretischen Perspektive zielt dabei auf das Unterrichtsgeschehen und die Lehrer-Schüler-Kommunikation, die neo-institutionelle Perspektive verknüpft das Unterrichtsgeschehen mit der administrativen Ebene. Anders zusammengefasst beschreibt ersteres das Verhältnis von Prozess-Dimension und Output-Dimension, zweites hingen das Verhältnis von Input-Dimension zu Prozess-Dimension (Diemer & Kuper, 2011). Dies soll nachfolgend erläutert werden.

Als Ausgangslage der Schulqualitätsforschung kann erstens ein hoher Allgemeinbildungsanspruch für alle und zweitens die Erkenntnis angenommen werden, dass die Besonderheit der situativen Umstände als entscheidend für die Wirkung von Schule angesehen werden kann (Fend, 2009; Rolff, 1993). Entsprechend erhält die Einzelschule eine herrausgehobene Stellung. Ihr obliegt es, ein an Schüler und Schülerinnen und Schulumfeld adaptiertes Lernangebot vorzuhalten und den staatlichen Bildungsauftrag im Sinne des Subsidiaritätsprinzips zu erfüllen. Maag Merki (2010) zieht hier die Parallele vom Angebot-Nutzungs-Modell von Fend (2008) zur Educational Governance-Perspektive, da auch dort eine „Kundenorientierung“ vorgesehen ist. Allerdings sind in Fends Angebot-Nutzungs-Modell die Kunden nicht der Staat oder die Schüler und Schülerinnen bzw. Erziehungsberechtigte, sondern den Lehrkräften werden Informationen über den Lernstand ihrer Schülerinnen und Schüler angeboten, die sie – das Verständnis der Informationen vorausgesetzt − für die Verbesserung der Unterrichtsqualität nutzen können.

Die Instrumente der „Neuen Steuerung“ können in zweifacher Art als Umgang mit einem Technologiedefizit verstanden werden: indem der Blickwechsel auf das Ergebnis bzw. das Ziel von Unterricht gelenkt wird (vgl. auch implementierende und deskriptive Funktion) und indem Steuerungswissen durch die Darstellung des Ergebnisses zugänglich gemacht wird (vgl. u.a. diagnostische Funktion). Dabei wird vorausgesetzt, dass Unterrichten stets eine Art Problemlöseprozess darstellt, bei dem Schülerinnen und Schüler mit Aufgaben konfrontiert werden und die Lehrkraft versucht, ihnen Lösungsstrategien zu vermitteln, um bei den Schülern und Schülerinnen die Entwicklung von Problemlöseprozessen wahrscheinlicher werden zu lassen. Einerseits kann nun angenommen werden, dass eine Output-Orientierung zu einer Anpassung der Prozesse (Aufgabenbearbeitung durch die Schülerinnen und Schüler und Unterstützung durch die Lehrkraft) im Vorfeld des Outputs (Anwenden der Problemlöseprozesse) führt. Das entspricht einem Feed-forward und ist eine der Intentionen des Steuerungswechsels (Heymann & Pallack, 2007). Das Technologiedefizit wird aber nur dann tatsächlich vermindert, wenn das durch die Evaluation angebotene Wissen auch zu

[18] Das „Technologiedefizit der Pädagogik“ meint im Sinne Luhmanns die Schwierigkeit zu unterrichten, ohne dabei auf Kausalitätsannahmen zurückgreifen zu können. Kausalitätsannahmen kann es aber nicht geben, da die zu Unterrichtenden selbstreferentielle Bewusstseinssysteme sind (Fend, 2009).

[19] Für eine abstraktere Betrachtung siehe Maier (2009).

einer Revision von subjektiven Kausalplänen der Lehrkräfte führt. Dies wird durch zwei Faktoren erschwert: Neben den selbst steuerbaren Prozessen (beispielsweise der Unterrichtsgestaltung) können erstens auch diejenigen Prozesse hinterfragt werden, die man nicht selbst beeinflussen kann (beispielsweise die Zusammensetzung der Schülerschaft). Daher wird angenommen, dass beispielsweise „Faire Vergleiche“[20] bei standardisierter Schulleistungsmessung nötig sind, um den Blick auf die steuerbaren Prozesse lenken zu können.[21] Anderenfalls bietet die Beschäftigung mit nicht selbst beeinflussbaren Prozessen zu viel Raum für ausweichende Analysen der gewonnenen Daten, indem die Aussagekraft der Ergebnisse bezweifelt wird. Zweitens ist Evaluationswissen erst einmal nur Beobachterwissen und muss von den Lehrkräften in Handlungsalternativen transformiert werden (Diemer & Kuper, 2011). Das größte Problem bleibt auch innerhalb dieses Ansatzes die Frage, ob Lehrkräfte Handlungsalternativen bzw. Prozessalternativen kennen, um bessere Ergebnisse erzielen zu können. Unterstellt man nicht, dass Lehrkräfte mit schlechteren Ergebnissen lediglich die falschen Prozessstrategien vermittelten bzw. anwendeten und auf die falschen Ziele ausgerichtet handelten, muss weniger in die Generierung von Produktwissen (Wissen über den Unterrichtserfolg) als mehr in Weiterbildung investiert werden. Zumindest müssen Maßnahmen, mit denen Produktwissen gewonnen wird, mit Maßnahmen flankiert werden, die Handlungsalternativen aufzeigen.

Die neo-institutionelle Perspektive (Weick, 1976) gehört zu den bürokratietheoretischen Ansätzen. Berkemeyer (2010) hält es nach wie vor für zulässig, die Mehrebenenstruktur des Schulsystems derart zu beschreiben, da die klassische Erlassstruktur ein „Top-down-Phänomen“ darstellt. Pädagogische Prozesse sollen erkennbar extern geregelt werden (Berkemeyer, 2010). Die neo-institutionelle Perspektive betrachtet die zu übergeordneten Hierarchieebenen in loser Kopplung agierende Einzelschule (Kuper, 2001). Die lose Kopplung wird dabei ursprünglich erst einmal wertfrei als Voraussetzung für Flexibilität und Funktionalität gesehen, wodurch wiederum ein funktionales Verhältnis von der technischen Umwelt (Anforderung zu unterrichten) und institutioneller Umwelt (Vorgaben, Regelungen) ermöglicht wird (Maier, 2009). Lehrkräfte sehen in den Erlassen übergeordneter Hierarchieebenen eine Gefährdung ihrer Kernaufgabe (zu unterrichten). Sie verschaffen sich die in ihren Augen für das Unterrichten notwendige Freiheit, indem externe Forderungen nur symbolisch (d.h. nur scheinbar und oberflächlich) adaptiert werden und möglicherweise keinen Einfluss auf die Kerntechnologie haben. Im Zusammenhang mit der Neuen Steuerung bedeutet eine lose Kopplung nur eine schwache Korrelation von Entwicklung neuer Vorgaben i.S. der Bildungsstandards und neuen Lernplänen als Input-Dimension und ihrer Implementation in Schule und Unterricht als Prozess-Dimension (Diemer & Kuper, 2011). Testbasierte Schulreformen sollen nun die Kopplung zwischen Einzelschule und

[20] „Faire Vergleiche“ bezeichnen rückgemeldete Vergleichsdaten von anderen Testteilnehmenden, die ähnliche Kontextvoraussetzungen (z.B. Zusammensetzung der Schülerschaft) aufweisen könne. Ein alternatives Verfahren ist die Adjustierung, bei der aus den realen Testwerten unter Berücksichtigung von unterschiedlichen Kontextvoraussetzungen neue Testwerte ermittelt und zurückgemeldet werden.

[21] „Faire Vergleiche“ reichen allerdings für eine umfangreiche Nutzung der Ergebnisse nicht. Dies zeigen auch Analysen von Maier, Metz, Bohl, Kleinknecht und Schymala (2012).

Bildungsadministration neu regeln, indem inhaltliche Vorgaben reduziert werden. Darin liegt die Hoffnung, dass Anforderungen der institutionellen Umwelt nicht einfach „ausgesessen" werden, sondern tatsächlich auf die technische Umwelt wirken können. Durch die Verknüpfung institutioneller Vorgaben mit Tests soll eine direkte Verbindung zum Unterrichten hergestellt werden.

Problematisch hieran ist, dass eine hohe Selbstständigkeit der Einzelelemente eine Bedingung für eine notwendige Adaption an neue Umweltbedingungen darstellt (Terhart, 1986) und dass eine engere Kopplung zwischen der Schule und einzelnen Lehrkräften oder zwischen der Schule und der übergeordneten Hierarchieebene nur dann zu mehr Qualität führt, wenn dies mit einer besseren pädagogischen Ausgestaltung der Reform einhergeht (Maier, 2009) (vgl. Diemer oben). Typische Fehlwirkungen dieser engen Kopplung sind im nächsten Abschnitt (2.1.2) dargestellt. Empirisch haben sich stattdessen die Akzeptanz und das Verständnis der neuen Verfahren sowie die Berücksichtigung der situativen Ausgangslage als notwendige Voraussetzungen für den Erfolg von testbasierter Reformen gezeigt (dazu für Deutschland (2.4) und (4.3.5)).

Die Perspektive der **Schulentwicklungsforschung** erweitert die Analyse der Wirkungsweise jener Instrumente der Neuen Steuerung um die nötigen schulischen Prozesse, durch die mit den beschriebenen Anforderungen und Zielen umgegangen werden kann (Maag Merki, 2010). Zentral ist hierbei der Gedanke von Schule als lernende und autonom agierende Institution. Als Instrument dieses angenommenen Lernprozesses wird Lernen durch Evaluation gedacht und entsprechend werden Modelle zu Evaluationsprozessen herangezogen. Als Beispiel für den deutschsprachigen Raum können das vierphasige Rahmenmodell zum pädagogischen Nutzung von Ergebnissen aus Vergleichsarbeiten von Helmke und Hosenfeld und das fünfstufige Modell zur Steigerung der Unterrichtsqualität von Tresch dienen (Helmke & Hosenfeld, 2005; Tresch, 2007)[22]. Beide Modelle sind allerdings lineare und eindimensionale Modelle. Sie entsprechen daher eher der Tradition der Schuleffektivitätsforschung (als deren Vertreter Helmke und Hosenfeld gelten können) und bilden die komplexen Wirkungen innerhalb einer Schule nur unzureichend ab (Maier, 2009). Im Prinzip beschreiben sie den idealen Verarbeitungsprozess von Daten aus standardisierter Schulleistungsmessung durch eine einzelne Person. Fachgruppen oder Jahrgangsstufenteams werden dabei als Sammlung von exakt gleich handelnden Lehrkräften angesehen. Rolff sieht Schulentwicklung und testbasierte Schulreformen daher auch als zwei völlig verschiedene Qualitätsmanagementmodelle (Rolff, 2007). Die Schulentwicklungsforschung untersucht nicht, wie Schulen auf Anforderungen von außen reagieren, sondern sieht die Einzelschule als Gestaltungseinheit. Dementsprechend gibt sie sich auch nicht mit der individuellen Perspektive von Implementations- und Innovationsprozessen zufrieden, sondern betrachtet das Zusammenwirken der Individuen innerhalb einer Schule. Typische Forschungsschwerpunkte sind das Unterstützungsverhalten des Rezeptions- und Reflexionsprozesses durch die Schulleitung, die Bedeutung professioneller

[22] Siehe Abschnitt (2.2.4).

Lerngemeinschaften und anderer Kooperationsformen für die Wirkung der Neuen Steuerung, Wirkungen auf die Schulprogrammarbeit oder die Effekte externer Unterstützung.

Mit der Skizze der historischen Entwicklung von SSL, der drei wichtigen wissenschaftlichen Perspektiven und den im vergangenen Abschnitt dargestellten Zielen ist ein steuerungstheoretischer Rahmen im Sinne einer deskriptiven Betrachtung gesteckt. Offen bleibt bisher die Frage nach der Wirksamkeit[23] von SSL, um die Ziele zu verwirklichen. Wie bereits u.a. in der Einleitung zu diesem Kapitel beschrieben lassen sich zwei Steuerungsmechanismen ausmachen: Rechenschaftslegung und Unterstützungsangebote bei der Qualitätsentwicklung. Diese beiden Steuerungsmechanismen werden im nächsten bzw. übernächsten Abschnitt dargestellt.

2.1.2 Wirkungen von standardisierter Schulleistungsmessung als Rechenschaftslegung

Die Wirksamkeit von SSL als Instrument der schulischen Rechenschaftslegung wird im internationalen Kontext unter dem Begriff „Accountability“[24] diskutiert. Wirkungen auf der Makroebene, der Einzelschulebene und der Klassenebene werden auf Grundlage aggregierter Daten berichtet. Die Befunde zu den Wirkungen auf die Schülerleistung, zur Entwicklung des Unterrichts, aber auch über den Umgang mit dem Anspruch der Rechenschaftslegung sind komplex. Zusätzlich werden nicht-intendierte Wirkungen beschrieben, die die eigentlichen Intentionen unterlaufen und die Validität der eingesetzten Tests in Frage stellen. In welcher Weise eine Qualitätsentwicklung innerhalb von Schulen stattfindet, spielt in den Betrachtungen nur eine untergeordnete Rolle.

Wenn bisher von Rechenschaftslegung gesprochen wurde, wurden die verschiedenen Formen der Rechenschaftslegung nicht unterschieden. Für eine bessere Differenzierung der Rechenschaftslegung kann in der ersten Dimension unterschieden werden, wer Rechenschaft ablegt, und in der zweiten Dimension, wem gegenüber Rechenschaft abgelegt wird (Adressatendimension). Für die schulische Rechenschaftslegung kommen die Schülerinnen und Schüler, die Lehrkräfte, die Einzelschule als direkt für den Lernprozess Verantwortliche und die Bildungsadministration als Verantwortliche für die Unterstützungsleistung in Betracht (Linn, 2004). Auf der Adressatendimension kann man im schulischen Kontext von Rechenschaftslegung gegenüber der schulischen Administration (bürokratische R.) oder den Erziehungsberechtigten und Schülerinnen und Schülern bzw. den

[23] Unterschieden werden „Wirkungen“ und Wirksamkeit derart, dass unter „Wirksamkeit“ erfüllte Intentionen, unter „Wirkung“ hingegen alle beobachteten Effekte gefasst werden Lamprecht und Rürup (2012).

[24] Ausgangslage dieser Betrachtung ist der englische Begriff „Accountability“, der mehr meint als nur Rechenschaftslegung. Weitere Formen können daher auch „legal Accountability“ und „moral Accountability“ sein Crundwell (2005). „Legal Accountability“ ist im schulischen Zusammenhang dasgleiche wie bürokratische Rechenschaftslegung Darling-Hammond (2004).

Bürgerinnen und Bürgern (politische R. bzw. marktorientierte R.) sprechen. Nach Anderson handelt das sich als rechenschaftspflichtig wahrnehmende Subjekt im ersten Fall nach Vorschriften, im zweiten Fall auf den Lernerfolg der Schüler und Schülerinnen ausgerichtet. Neben diesen beiden Formen kann auch noch von Rechenschaftslegung innerhalb einer Profession gesprochen werden. Die Beteiligten orientieren sich an Standards der Profession wie beispielsweise den Standards der Deutschen Gesellschaft für Evaluation. Die Beteiligten fühlen sich dann gegenüber der Professionsgemeinschaft rechenschaftspflichtig (Anderson, 2005). Nach O`Day handelt es sich bei schulischer Rechenschaftslegung um eine Kombination aus bürokratischer und aus professioneller Rechenschaftslegung (O`Day, 2004).

Für Anderson baut ein gutes Rechenschaftssystem auf fünf Punkten auf: (1) klaren Zielvorstellungen, (2) Tests zur Überprüfung, ob die Ziele realisiert wurden, (3) an die Herausforderungen angepassten Anweisungen, (4) Unterstützungsmaßnahmen zur Datennutzung und (5) positiven wie negativen Sanktionen (Anderson, 2005).

Die Funktionslogik schulischer Rechenschaftslegung (von Fuhrman mit Verweis auf Argyris und Schön als „Theory of Action“ bezeichnet) im Zusammenspiel mit Standards und SSL basiert auf vier Schritten: (a) Der Lernerfolg der Schülerinnen und Schüler ist das Ziel der schulischen Ausbildung und SSL betont den Wert von guten Schülerleistungen. (b) Die eingesetzten Instrumente der SSL können diese Schülerleistung messen. (c) Konsequenzen, die auf die gezeigten Schülerleistungen folgen, motivieren die Beteiligten.[25] (d) Eine höhere Motivation führt zu Qualitätsentwicklung und besseren Leistungen (Fuhrman, 2004). Van Ackeren bezeichnet dieses Wirkungsmechanismus als „Pressure-Ansatz“. Der Ansatz galt lange als typisch für den Bildungsbereich im anglo-amerikanischen Raum (van Ackeren, 2003). Grundsätzlich wird dabei ein behavioristisches Verständnis zugrunde gelegt (Ryan & Sapp, 2005). Dieses betrachtet die einzelne Lehrkraft, aber auch die Einzelschule als Einheit, die ihr Handeln von positiven wie negativen Sanktionen leiten lässt. O`Day weist auf die Unterkomplexität dieses Ansatzes hin. Wird von Schulen Rechenschaft verlangt, bleib es trotzdem den einzelnen Individuen (Lehrkräften) überlassen, in gewünschter Weise zu handeln. Dies stellt in der Beschreibung der Wirkungsweise dieser Art von Steuerung neben der externen Kontrolle interner Prozesse (s.a. Principal-Agent-Theorie) sowie der Informationsnutzung das größte Problem dar (O`Day, 2004).

Der SSL kommt in diesem Konstrukt folglich eine zentrale Bedeutung zu. Diese konkretisiert sich in Form von so genannten „High-Stake-Tests“. Als High-Stake-Tests werden Tests bezeichnet, deren Ergebnis eine wegweisende Bedeutung für die jeweiligen Testpersonen besitzt, wohingegen dies bei Low-Stake-Tests nicht angenommen werden kann. Typische Sanktionen bei High-Stake-Tests sind nach Ryan und Brown finanzielle Anreize und

[25] Jürgens und Schneider halten nach Analysen von PISA- und TIMSS-Daten ein anderes, einfachere Modell für geeigneter, um Leistungssteigerungen bzw. den Leistungsvorsprung von Staaten mit standardisierten Schulleistungsmessungen gegenüber Staaten ohne diese Prüfungen zu erklären. Da von Schülerinnen und Schülern in den leistungstärkeren Staaten kein „besserer Unterricht“ berichtet wird, führen sie die höheren Testleistungen vorwiegend auf eine größere Anstrengung der Schüler und Schülerinnen zurück (Jürges und Schneider, 2008).

Arbeitsplatzgarantien für Lehrkräfte bzw. Schulen sowie Klassenwiederholung oder Zertifizierung und Versetzung für Schüler und Schülerinnen (Ryan & Brown, 2007)[26]. In der Hälfte der Fälle sind in Europa und Nordamerika auch diejenigen zentralen Lernstandserhebungen als High-Stake-Tests angelegt, die der Qualitätsentwicklung dienen sollen (Parveva et al., 2009)[27]. Die Einführung von High Stake-Tests als Steuerungsinstrument ist umfangreich im anglo-amerikanischen Raum untersucht worden und soll nachfolgend kurz skizziert werden.

Die Wirkungen von standardisierter Schulleistungsmessung auf die Schülerleistungen

Das eigentliche Ziel von High-Stake-Tests ist ein gesteigerter Kompetenzerwerb (u.a. die Sicherung von Mindeststandards) bei den Schülerinnen und Schülern durch testbasierte Schulreformen. Ob dieses Ziel erreicht wird, ist unter Forschern umstritten. Je nach Stichprobe und Klassifikation des Grades der Rechenschaftslegung kann es zu sehr unterschiedlichen Befunden kommen. Die Diskrepanz zwischen den einzelnen Studien liegt wahrscheinlich an der zu komplexen Struktur der Einflussgrößen auf die Schülerleistung (Hanushek & Raymond, 2006). Dies zwingt die Forscher beispielsweise die innerschulischen Prozesse auszublenden (Maier, 2010b). Ein typisches Beispiel sind die Befunde von Amrein und Berliner einerseits und die Befunde Rosenshines bzw. von Raymond und Hanushek andererseits (Amrein & Berliner, 2002, 2003; Raymond & Hanushek, 2003; Rosenshine, 2003). Die Studien verglichen die durchschnittlichen Schülerergebnisse verschiedener Bundesstaaten der U.S.A. Dabei dienen die jeweilige Ergebnisse aus zentralen Abschlusstests wie dem Scholastic Achievement Test (SAT)[28] und aus zentralen Lernstandserhebungen wie dem US-weiten National Assessment of Educational Progress **(**NAEP)[29] als Referenzwert. Amrein und Berliner verglichen Bundesstaaten, die High-Stake-Tests eingeführt hatten, und

[26] Maag Merki, Holmeier, Jäger und Oerke bezeichnen das Zentralabitur in Deutschland als „Low-Stake-Test“, da die Ergebnisse nur für die Schüler und Schülerinnen Relevanz besäßen (Maag Merki, Holmeier, Jäger & Oerke, 2010). Diese Einschätzung ist mit Blick auf das Professionsverständnis von Lehrkräften zu hinterfragen, entspricht aber auch nicht dem gängigen Verständnis von High-Stake-Tests (vgl. auch Heubert, 2004). Brunner und Kollegen binden den Begriff „High-Stake“ ausschließlich an die Relevanz für Schüler und Schülerinnen (Brunner, Artelt, Krauss &Baumert, 2007). Van Ackeren, Block, Klein und Kühn unterscheiden zwischen Abschlussprüfungen mit „High-Stakes“ für Lehrkräfte und Schüler sowie „High Stakes“ nur für Schüler (Ackeren, Block, Klein & Kühn, 2012). Lind unterscheidet bei den mit standardisierter Schulleistungsmessung verbundenen Zielen zwischen Instrumenten zur Personenevaluation (oder auch Personenbeurteilung) und Instrumenten zur Programmevaluation. High-Stake-Tests dienen grundsätzlich der Personenevaluation, da Verbesserungswirkungen von den evaluierten Beteiligten nicht aus sich heraus angestrebt werden, sondern das Resultat von persönlich erfahrenen Sanktionen sind (Lind, 2009). Im Folgenden wird auf eine Definition verzichtet und im nächsten Abschnitt eine Einordnung von zentralen Vergleichsarbeiten in Deutschland ohne diese Kennzeichnung vorgenommen.

[27] Für formative Zwecke werden nur ca. ein Drittel der zentralen Lernstandserhebungen genutzt (ebenda).

[28] Der SAT wird ähnlich dem deutschen N.-C.-Verfahren in den U.S.A genutzt, um die Hochschulzulassung zu regeln.

[29] Der NAEP ist eine jährlich in den Stufen 4, 8 und 12 in allen US-Bundesstaaten durchgeführte zentrale Lernstandserhebung auf Stichprobenbasis. Die Stichprobe ist für die Erhebung in den Stufen 4 und 8 groß genug, um jeweils Unterschiede zwischen den Bundesstaaten, verschiedenen Ethnien und dem Geschlecht für die Testbereiche Lesen, Schreiben, Mathematik und Naturwissenschaften vergleichen zu können. Es werden keine Schul- oder Individualergebnisse veröffentlicht.

setzten die Testwerte vor und nach der Einführung in Beziehung. Für ihre Analyse haben sie die 18 Bundesstaaten verglichen, die ihrer Ansicht nach den größten Druck mit SSL ausüben. Dabei konnten sie über alle herangezogenen Referenzgrößen jeweils mehr Staaten ermitteln, die nach der Einführung von High-Stake-Verfahren schlechtere Testleistungen zeigten als der US-weite Schnitt. Einige Staaten hatten sich allerdings auch verbessert (Amrein & Berliner, 2002). Raymond und Hanushek hingegen kritisieren die Arbeit von Amrein und Berliner als methodisch unsauber. Nachvollziehbar bemerken sie, dass Amrein und Berliner Schülerleistungen von jeweils Viertklässlern bzw. Achtklässlern vergleichen, die unterschiedlichen Kohorten angehören, statt den Lernzuwachs der ersten Kohorte zu ermitteln. Ob dabei aber ausgerechnet die Einführung von hoher Rechenschaftslegung bzw. High-Stake-Tests die entscheidende Größe darstellt, muss bezweifelt werden. Ihren Untersuchungen nach bringen Schulsysteme mit hohem Grad der Rechenschaftslegung die besten Schülerleistungen hervor, welches sie durch größere Zuwachsraten bei Schülerleistungen in Staaten mit entsprechendem Grad der Rechenschaftslegung belegt sehen. Vor allem die Schülerinnen und Schüler aus Staaten ohne jegliche Rechenschaftslegung haben sich in ihren Untersuchungen im Verhältnis viel geringer verbessert (Raymond & Hanushek, 2003). Auch Rosenshine kommt zu ähnlichen Ergebnissen, wenn er die von Amrein und Berliner ausgewählten Staaten mit Staaten ohne Rechenschaftslegung vergleicht (Rosenshine, 2003). Dabei berichtet er wie auch Raymond und Hanushek allerdings nur einen Durchschnittswerts über alle Staaten, sodass ihre Ergebnisse ebenfalls mit Vorsicht zu betrachten sind. Auch Carnoy und Loeb berichten über einen positiven Effekt von standardisierten High School-Examen und Testwertzunahmen (Carnoy & Loeb, 2004).

Nach einem ähnlichen Muster kommen Baumert und Watermann (2000) zu dem Ergebnis, dass zentrale Abschlussprüfungen teilweise der Schülerleistung förderlicher sein könnten und bezüglich eines Mindestniveaus normierend wirken könnten.[30] Sie verglichen die Ergebnisse der sechzehn Bundesländer in Deutschland anhand der Punktwerte für Mathematik und Physik im Rahmen der TIMSS-Oberstufen-Studie 1995. Länder mit Zentralabitur erreichten in Mathematik in Grundkursen durchschnittlich tendenziell etwas bessere Werte und weisen eine geringere Leistungsstreuung auf. Allerdings zeigte sich dieses Ergebnis nicht für das Fach Physik (Baumert & Watermann, 2000).

Fuchs und Wößmann fanden bei einer Reanalyse von Daten aus TIMSS 1995, TIMSS 1999, PISA 2000 und PISA 2003 zumindest für Mathematik und Naturwissenschaften signifikant bessere Ergebnisse in Staaten mit standardisierten Abschlussprüfungen als in Staaten, die keine standardisierten Abschlussprüfungen durchführen (Fuchs & Wößmann, 2007; Wößmann, 2008)[31]. Durch standardisierte Abschlussprüfungen scheinen außerdem

[30] Man beachte den Unterschied im Kriterium zwischen den US-amerikanischen und beiden folgenden Studien: Im ersten Fall geht es um Rechenschaftslegung, im zweiten um Standardisierung durch standardisierte Abschlussprüfungen.

[31] Die Aussagekraft dieser Untersuchung wird allerdings dadurch eingeschränkt, dass globale Untersuchungen das tatsächliche Bild häufig nicht richtig abbilden. Gründe dafür sind einmal unterschiedliche Strukturen in föderal angelegten Staaten und falsche Angaben durch die befragten Personen auf administrativer Ebene. Auch

Autonomiekonzepte für Einzelschulen effektiver zu werden. Entscheidend scheint zu sein, dass durch standardisierte Abschlussprüfungen klarere Zielvorgaben gegeben werden (Wößmann, 2007). Diese Interpretation der Datenlage ist allerdings nicht unumstritten, wie beispielsweise Block, Klein, Ackeren und Kühn (2011) schreiben.

Neben dem Ziel, die durchschnittliche Schülerleistung insgesamt zu steigern, soll durch SSL auch die soziale bzw. ethnische Selektion durch Schulen vermindert werden. Tatsächlich zeigt sich aber beispielsweise in verschiedenen Studien ein konstanter Unterschied zwischen Schülerinnen und Schülern verschiedener Bevölkerungsgruppen in den USA. Schüler und Schülerinnen ethnischer Minderheiten und Schülerinnen und Schüler aus sozioökomischschwächeren Schichten erreichen geringere Testwerte (Burns, Courtad, Hoffman & Folger, 2004). Insbesondere kann die Lücke zwischen weißen amerikanischen Schülerinnen und Schülern und afro-amerikanischen nicht geschlossen werden. Während sie sich zwischen Weißen und Migranten aus dem mittelamerikanischen Gebiet reduziert, weitet sich der Abstand zwischen Weißen und Schwarzen durch die Einführung von „Accountability-Systemen“ sogar aus (Hanushek & Raymond, 2004, 2006).[32] Als Grund wird die häufige Praxis der Bildungspolitik angesehen, einen höheren Grad der Rechenschaftslegung nicht mit größeren Unterstützungsmaßnahmen zu verbinden, sondern diese durch die Rechenschaftslegung scheinbar zu ersetzen (Lee & Wong, 2004; Lee, 2006, nach Maier, 2010). Als weitere Erklärungen dienen die eher an Schichten mit höherem sozioökonomischen Status orientierte Testsprache und die Strategie von Schulen, sich ausschließlich um die Förderung derjenigen Schüler und Schülerinnen zu kümmern, bei denen größere Steigerungen in Bezug auf den Testwert zu erwarten sind (Jones, 2007). Eher ist eine Zunahme der sozialen Selektion zu erwarten, wenn wie beispielsweise nach dem No-Child-Left-Behind-Gesetz gute Schulen auch für gute Lehrkräfte immer attraktiver werden. Eine Folge seien auch die steigende Quote der Schülerinnen und Schüler aus der Gruppe ethnischer Minderheiten, die vorzeitig die Schule abbrechen (Darling-Hammond, 2004). Carnoy hält dem allerdings entgegen, dass die Befunde zu hohen Abbrecher- und Sitzenbleiberquoten nicht konsistent sind und sich auch in Staaten ohne hohen Grad der Rechenschaftslegung finden lassen. Viel mehr ist nach seinen Untersuchungen der relative Grad der Ausbildung von spanischen Muttersprachlern in den U.S.A sogar höher als der von weißen Amerikanern, wenn man bei der Berechnung die Testleistungen aus zentralen Prüfungen kontrolliert (Carnoy, 2005).

konnten Block et al. die Ergebnisse auf Grundlage der PISA2003-Daten nicht replizieren. Die Unterschiede erwiesen sich ihren Analysen zufolge entweder als minimal oder zeigten sich für die einzelnen Bundesländer nicht in allen getesteten Bereichen (Block, Klein, Ackeren & Kühn, 2011; Block & van Ackeren nach Kühn, 2010).

[32] Hanushek und Raymond widersprechen damit Carnoy und Loeb, die keine strukturellen Benachteiligungen zwischen Ethnien fanden. In ihren Analysen zeigten sich grundsätzlich positive Effekte von Systemen der starken schulischen Rechenschaftslegung im Vergleich zu Staaten ohne Rechenschaftslegung (Carnoy & Loeb, 2004).

Die Wirkungen auf den vorgelagerten Unterricht

Man spricht auch von „Washback"-Effekten, wenn die Wirkungen von schulischer Rechenschaftslegung mittels Tests auf das Lernen und Unterrichten gemeint ist (Cheng & Curtis, 2004). Die Einführung von Verfahren der SSL kann zu einer Veränderung der Unterrichtspraxis in positiver wie in negativer Richtung führen, beispielsweise je nachdem, ob die Bindung des Curriculums an die Testinhalte zu einer Engführung oder zu einer Konkretisierung der Unterrichtsinhalte führt. Au berichtet in einer Meta-Analyse aus neunundvierzig qualitativen Studien über die Wirkungen von SSL auf den Unterrichtsinhalt und die Unterrichtsmethoden. In achtzig Prozent der untersuchten Studien waren solche Wirkungen nachgewiesen worden. In knapp dreißig Prozent der Fälle war das Curriculum erweitert worden, doppelt so häufig war aber eine Engführung auf testrelevante Inhalte berichtet worden. Auch hat sich in den qualitativen Studien häufiger eine Entwicklung zur Vermittlung von Wissenshäppchen als zu einer integrativen, verbindenden Wissensvermittlung gezeigt und der Unterricht entwickelte sich eher zu einem lehrer-zentrierten statt zu einem schüler-zentrierten Unterricht. Allerdings sei, so Au, dies keine prinzipielle Folge von High-Stake-Tests. Viel mehr kommt es offensichtlich auf die exakte Ausgestaltung der Tests an, sodass auch schüler-zentrierterer Unterricht erreicht und eine inhaltliche Engführung vermieden werden kann, wenn die Testaufgaben dies nicht begünstigen (Au, 2007). Es lässt sich schließen, dass die Ausgestaltung von den Verfahren der SSL und die Gestaltung der mit ihnen verknüpften Standards entscheidend für den Erfolg dieser Art von Steuerung sind (Baker & Linn, 2004; Green, 2006).

Ähnliche Hinweise über eine curriculare Einschränkung berichten auch Abrams sowie Amrein und Berliner für US-amerikanische Schulen. Van Ackeren bilanziert auch für Schulen in England eine entsprechende Wirkung von standardisierten Tests auf das Schulcurriculum bzw. das unterrichtete Curriculum (2005). Wie Au berichten auch Abrams sowie Wegener, Fromme und Clausen über einen auf Prüfungen ausgerichteten lehrer-zentrierten Unterricht (Abrams & Madaus, 2003; Amrein & Berliner, 2002; Wegener, Fromme & Clausen, 2011). Dies entspricht den Befunden zur Selbstbestimmungstheorie der Motivation nach Deci und Ryan, nach der auf Kontrollmotiven basierende Motivation (beispielsweise bei externaler Verhaltenssteuerung oder starker Relevanz für den Selbstwert einer Person) zu weniger qualitätsvollem Lernen, weniger Ausdauer und mehr negativen Emotionen führt (Deci et al. 1982 nach Ryan & Brown, 2007, vgl. auch van Ackeren, Block, Klein & Kühn, 2012).

Entsprechende inhaltliche Verkürzungen werden dabei nicht nur inoffiziell von den Lehrkräften vorgenommen und berichtet (Stecher, Chun & Barron, 2004), sondern sind ein offizieller Bestandteil des Schulcurriculums. Diese beschränken sich auf mit den speziellen Testverfahren prüfbare Inhalte (Stecher & Barron, 2001). Vielmals zeigen Studien außerdem, dass an US-amerikanischen Schulen eine gezielte Vorbereitung auf Tests (sowohl auf High-Stake- als auch auf Low-Stake-Tests) stattfindet (siehe insbesondere Kap. 3). Dabei werden neben sinnvollen Kompetenzen auch Fertigkeiten vermittelt, die aus Wissen über die Tests resultieren und die Testvalidität reduzieren. Systemmonitoring wird dadurch genauso

unmöglich wie eine Rückmeldung auf Klassen- oder Individualebene (Herman, 2004; Koretz, 2008).

Maag Merki, Holmeier, Jäger und Oerke (2010) haben Wirkungen auf die Unterrichtsqualität durch die Einführung des Zentralabiturs in Bremen und Hessen untersucht. In ihrer vergleichenden Studie, die die Unterrichtsqualität der Leistungskurse in den Jahren 2007 (Einführung zentraler Prüfungen in Leistungs- und Grundkursen in Hessen, nur in Grundkursen in Bremen) und 2008 (Einführung von zentralen Prüfungen in den Leistungskursen Deutsch, Englisch, Mathematik und den Naturwissenschaften in Bremen) in den Fächern Biologie, Deutsch, Englisch und Mathematik mittels Schülerbefragungen untersuchte, zeigten sich positive Effekte auf das Ausmaß der kognitiven Aktivierung durch die Einführung von zentralen Abiturprüfungen in Leistungskursen in den Fächern Mathematik (nur Bremen) und Englisch (in beiden Ländern), aber negative Effekte in Englisch (in Hessen). Für die Fächer Deutsch und Biologie zeigten sich in den Leistungskursen keine Effekte. Maag Merki und Kolleginnen führen diese eher positiven Veränderungen der Unterrichtsqualität auf das Zusammenspiel zwischen zentralen Prüfungen und den gleichzeitigen geringeren Druck für Lehrkräfte zurück, den Abiturprüfungen im Vergleich zu High-Stake-Tests für die Lehrkräfte selbst hätten (Maag Merki & Holmeier, 2008; Maag Merki, Holmeier, Jäger & Oerke, 2010). Demgegenüber stehen die Ergebnisse der Lehrerbefragung dieses Projekts aus den Jahren 2007 bis 2009. Jäger (2012) berichtet, dass Lehrkräfte unter den Bedingungen zentraler Abiturprüfungen eher angeben, die Themenvarianz in ihren Kursen eingeschränkt zu haben. Dies gilt besonders für Hessen, in denen von Anfang an auch in den Leistungskursen zentral geprüft wurde. Außerdem wird andersherum von mehr Unterrichtsvarianz und größerer Orientierung an Schülerinteressen durch die Lehrkräfte berichtet, wenn sich diese kollektiv selbstwirksamer fühlen, niedrige Unsicherheit empfinden und zu curricularen Themen kooperieren (Jäger, 2012).

Die von den Lehrkräften berichteten Veränderungen lassen sich aber wohl nicht durch Veränderungen der Abituraufgaben erklären, obwohl gerade diese das scheinbare Steuerungselement des Zentralabiturs darstellen. Kühn konnte nur sehr geringe Unterschiede zwischen Abituraufgaben mit dezentralem Abitur und zentralem Abitur in den naturwissenschaftlichen Fächern ermitteln. Dies lässt den Schluss zu, dass sich die Unterrichtsmethoden entweder nicht in den Prüfungsaufgaben ausdrücken oder (so schließt Kühn) andere Elemente als die Dezentralisierung/Zentralisierung von Prüfungen den Unterricht beeinflussen (Kühn, 2010).

Nicht-intendierte Nebenwirkungen

Zu einer eher lückenhaften Wirksamkeit von SSL kommt eine Reihe von Studien, in denen nicht-intendierte Nebenwirkungen und Betrugseffekte gefunden werden. Dabei wird das Verhalten der Lehrkräfte und der schulischen Administration durch Campbells Law illustriert, nachdem mit dem Druck auch Betrugsversuche zunehmen (Koretz, 2008).

Mehrere Studien berichten, wie Lehrkräfte ihre Förderung auf diejenigen Fälle konzentrieren, bei denen relativ schnell Verbesserungen zu erwarten sind (so genannte „Bubble Kids“) (van Ackeren, 2005). Außerdem wird befürchtet, dass das Prinzip, Lernen mit extrinsischer Motivation (den Tests) zu verbinden, die intrinsische Motivation der Schülerinnen und Schüler für das Lernen reduziert (Jones, 2007; Ryan et al., 2007). Möglicherweise führen zentrale Abschlussprüfungen zu einem höheren Grad der Fremdattribution von Schulleistungen (Oerke, Maag Merki, Holmeier & Jäger, 2011).[33] Ebenfalls gibt es Indizien dafür, dass der Grad der Rechenschaftslegung und die Versetzungsquote negativ korrelieren. Ein höherer Grad der Rechenschaftslegung führt folglich nicht zu einem höheren Grad der schulischen Bildung (Carnoy, 2005). Viel mehr wirken zentrale Tests in den häufig als Kernfächer bezeichneten Unterrichtsfächern ggf. auch negativ auf das Bildungsniveau anderer schulischer Bereiche und Fächer.

Unter Betrugseffekte fallen beispielsweise Strategien, schlechtere Schüler und Schülerinnen nicht an den Tests teilnehmen zu lassen. Dies vermuten verschiedene Forscher für Schulen in US-Bundesstaaten mit starker High-Stake-Orientierung. Amrein und Berliner zeigen beispielsweise eine starke Abnahme der Testteilnahmequote dieser Schülerinnen und Schüler nach Einführung von High-Stake-Tests in den Bundesstaaten ihrer Untersuchung (Amrein & Berliner, 2002). Nichols und Berliner listen auf Lehrereben außerdem klassische Betrugshandlungen als Reaktion auf High-Stake-Tests wie Vorsagen oder das Arbeiten mit unerlaubten Hilfsmitteln auf (2007b). Auch werden Testhefte offensichtlich nachträglich manipuliert, wie eine Prüfung auf Radierungen ergab (Otterman, 2011). Lehrerinnen und Lehrer halten sich dabei aus Sicht von Nichols und Berliner aus zwei Motiven nicht an die Regeln der schulübergreifenden Tests: Erstens aus rein egoistischen Gründen, weil sie persönlich Sanktionen befürchten, zweitens aber auch, weil sie ihren Schülerinnen und Schülern helfen und deren Chancen und Motivation erhöhen wollen.[34] Den Anteil der Lehrerinnen und Lehrer, die vorwiegend aufgrund des ersten Motivs betrügen, geben Studien zwischen 3% und 15% an (Nichols & Berliner, 2007). Schließlich kommt es auch auf administrativer Ebene zu Versuchen, bessere Testergebnisse durch unerwünschte Methoden zu erreichen. Nichols und Berliner berichten von frisierten Schulabbrecherquoten, Schuleinzugsbereichen und vereinfachten Tests in US-amerikanischen Bundesstaaten (Nichols & Berliner, 2007a). Ähnliches ist auch in europäischen Staaten mehrfach beobachtet worden. Auch die Schwerpunkte von Fortbildungen werden in einigen Staaten durch Tests bestimmt (Parveva et al., 2009).

Darling-Hammond bilanziert mit Blick auf die nur teilweise verwirklichten Ziele und nicht intendierten Nebenwirkungen von Steuerung durch SSL drei Herausforderungen für die Bildungsadministration, wenn Steuerung durch Rechenschaftslegung funktionieren soll: (1) Lehrkräfte müssen in die Lage versetzt werden, dass sie ausreichend Kompetenzen besitzen,

[33] Stecher weist auf die Schwierigkeit hin, solche Veränderungen zu messen, da es keine standardisierten Skalen gibt und nicht dieselbe Gruppe als Vergleichsgröße dienen kann (Stecher, 2002).

[34] Eine alternative Begründung liefern die Betrachtungen zu Widerstand als Reaktion auf Veränderungen (s. Abs. (4.3.1)).

um die Standards zu unterrichten. (2) Die organisatorischen Voraussetzungen an Schulen müssen eine hohe Unterrichts- und Lernqualität ermöglichen. (3) Die eingesetzten Evaluationsinstrumente müssen die tatsächlichen Lerngelegenheiten für Schüler und Schülerinnen und die Verbesserungseffekte erfassen, und zwar langfristig (Darling-Hammond, 2004). Diese Forderungen gelten aber ohne Abstriche auch für den Ansatz der datengestützten Qualitätsentwicklung durch standardisierte Schulleistungsmessung.

2.1.3 Datengestützte Qualitätsentwicklung durch standardisierte Schulleistungsmessung

Die alternative Steuerungsidee zu standardisierter Schulleistungsmessung ist die datengestützte Qualitätsentwicklung. Anders als im zuerst beschriebenen „Pressure"-Ansatz soll durch die SSL weniger das Engagement gesteigert werden, sondern das Augenmerk liegt auf der Nutzung der aus SSL gewonnenen Daten, um Systemmonitoring zu betreiben und Entwicklungsprozesse an den Einzelschulen anzuregen und zu unterstützen (Altrichter & Heinrich, 2006). Die grundlegende Vorstellung an dieser Stelle ist, dass Schulen und Lehrkräfte die Daten als Hilfeleistung erkennen und Schul- und Unterrichtsentwicklung aufgrund ihres Professionsverständnisses betreiben (Böttger-Beer & Koch, 2008; Visscher, 2008). Das durch SSL gewonnene Produktwissen kann für eine Evaluation genutzt werden.

Im Sinne von Evaluation als Prozess kann das Systemmonitoring durch SSL grundsätzlich als externe Evaluation bezeichnet werden (Berkemeyer & Müller, 2010). Wird die Interpretation und die Ableitung von Konsequenzen und Handlungsplänen der Einzelschule überlassen, handelt es sich um eine Mischform aus interner und externer Evaluation (sofern Fachgruppen diese Aufgaben übernehmen) oder um eine Mischform aus Selbst- und externer Evaluation (wenn nur die einzelne Lehrkraft mit dieser Aufgabe betraut ist).

Der Auftrag, aus den Daten rational Maßnahmen abzuleiten, lässt nach van Ackeren aber verschiedene Nutzungsformen zu (Johnson spricht von „Evaluation Utilization" Johnson, 1998): (a) Die *instrumentelle Nutzung* entspricht der eigentlichen Idee von Evaluation. Die gewonnenen Daten dienen direkt als Ausgangspunkt für abgeleitete Maßnahmen. Die Annahme ist, dass die Modifikationen von Programmen oder eine Umverteilung von Ressourcen auf Grundlage von gesammelten Daten und resultierenden Kausalzusammenhängen vorgenommen werden. (b) Die *konzeptionelle Nutzung* betrachtet die Ergebnisse und erhobenen Kausalzusammenhänge nur als eine Quelle für rationales Handeln und betrachtet mehrere Informationsfaktoren über einen längeren Zeitraum. (c) Von *symbolischer Nutzung* kann gesprochen werden, wenn die Daten lediglich taktisch genutzt werden, um politische Entscheidungen (möglicherweise nachträglich) zu legitimieren (Stockmann, 2006). (d) Die *prozesshafte Nutzung* findet auf der Metaebene statt. Anders als bei der instrumentellen und der konzeptionellen Nutzung liegt der Fokus hier nicht auf den konkreten Daten, sondern auf dem Prozess der Erhebung selbst. Durch die Partizipation der Erhebungsbetroffenen sollen ihnen die Ideen und Verfahren der Evidenzbasierung nahegebracht werden (van Ackeren, 2003).

Die Schulentwicklungsforschung auf Einzelschulebene und die Forschung zur pädagogischen Professionalität (vgl. die Abs. (2.2.4) und vor allem (4.3)) beschäftigen sich mit der Frage, wie in Schulen eine instrumentelle Nutzung[35] erreicht werden kann. Aus den Befunden der Schulentwicklungsforschung resultieren Modelle zu so genannten „School Performance Feedback Systems (SPFS)“, die die Faktoren untersuchen, unter denen eine datengestützte Qualitätsentwicklung in Schulen gelingt. Der Begriff geht auf Visscher[36] und Coe zurück (Visscher & Coe, 2003) und dient beispielsweise als Vorlage für das Modell zu SPFS von Verhaegh, Vanhoof, Valcke und van Petegem, welches gleichsam aus einer Zusammenstellung von früheren Befunden aus der Schulentwicklungs- und Feedbackforschung und aus qualitativen Interviews mit Grundschulschulleitungen von flämischen Schulen entwickelt wurde (Verhaeghe, Vanhoof, Valcke & Petegem, 2010).

Verhaegh und Kollegen unterteilen ihr SPFS in die Bereiche (A) beeinflussende Faktoren (Kontextfaktoren, Nutzungsressourcen, wahrgenommene Feedbackgüte, Unterstützungsfaktoren) (B) Feedbacknutzung (Evaluationszyklus, Nutzungsformen) und (C) Effekte. Das Modell berücksichtigt explizit die Wahrnehmung der Lehrkräfte bzgl. der Datengüte, des Veränderungsanlasses und den Unterstützungsbedarf wie auch den Kontext der Qualitätsentwicklungsmaßnahme als beeinflussende Faktoren. Dadurch wird die besondere Bedeutung der Beteiligten für eine gelungene datengestützte Qualitätsentwicklung innerhalb einer Einzelschule unterstrichen. Auch Erkenntnisse der Rezeptionsforschung finden indirekt Einzug in das Modell, indem die Nutzerfreundlichkeit integriert wird. Der Modellteil zur tatsächlichen Datennutzung mit sieben Phasen ist allerdings eher als Platzhalter zu verstehen.[37] Verhaegh u.a. (2010) selbst konnten die sieben Phasen bei nur sechzehn Interviews nicht alle abbilden. Ein Großteil der Schulen kam über die erste Phase nicht hinaus. An einigen Schulen wurden Lehrkräfte überhaupt nicht an der Rezeption der Daten aus der SSL beteiligt, da ihnen die Schulleiter diesen Aufwand nicht zumuten wollten und das Verstehen der Daten für zu schwierig hielten.

Ingram, Seashore Louis und Schroeder haben umgekehrt in einer Studie an neun US-amerikanischen High Schools Ergebnisse über die Entscheidungsfindung an Einzelschulen gewonnen. Sie formulierten sieben Barrieren, die unabhängig vom Rechenschaftsgrad wirken, wenn Einzelschulen und Lehrkräfte Daten aus SSL nicht nutzen, um dadurch Entscheidungen zu treffen. Hier sind zusätzlich zum SPFS von Verhaegh der Rückgriff auf (den Daten häufig widersprechende) Erfahrungen und der Wunsch nach einer Outcome- statt einer Output-Orientierung zu nennen (Ingram, Seashore Louis & Schroeder, 2004). Es wird für viele nicht ersichtlich, warum die „Qualität von Schule“ nur an der Beherrschung des

[35] Die Gestaltung der Tests und Besonderheiten der getesteten Stichprobe innerhalb einer Schule sprechen eher dafür, Daten aus einer SSL konzeptionell für die Qualitätsentwicklung zu nutzen. Mehrere Jahre umfassende Erhebungszeiträume sehen die Studien in der Regel aber nicht vor, sodass nur die instrumentelle Nutzung untersucht werden kann (z.B. Groß Ophoff, 2013; Koch, 2011; Verhaegh et al., 2010 – eine Ausnahme stellen Schildkamp, Visscher und Luyten, 2009, dar).

[36] Der Begriff SPFS wird konkreter als Bezeichnung für Qualitätsmanagementsysteme verwendet (z.B. ZEBO - Zelfevaluatie Basisonderwijs in den Niederlanden), die ein theoretisches Modell als Hintergrund haben (Visscher, 2008). Im Folgenden wird SPFS allerdings als Abkürzung für die Modelle benutzt, nicht die Systeme.

[37] Siehe Abb. (2.2.4) für alternative Modelle.

Schulstoffs in bestimmten Fächern gemessen wird statt den beruflichen Erfolg der Schülerinnen und Schüler zu berücksichtigen (Dubs, 2006).

Insgesamt zeigt die Befundlage von Verhaegh u.a., dass in Schulen nur selten datengestützte Schul- und Unterrichtsentwicklung stattfindet. Verhaegh u.a. fanden in ihrer qualitativen Studie nur wenige Belege für einen instrumentellen Datengebrauch. Beispielsweise gaben einige Schulleitungen als Schulentwicklungsmaßnahe an, Dienstpläne geändert und die Zahl der Unterrichtsstunden erhöht zu haben. Auch wurde stellenweise ein Mentoring für neue Lehrkräfte eingeführt. Als Maßnahme der Unterrichtsentwicklung kann die Implementation einer neuen Lesemethode angesehen werden (Verhaeghe et al., 2010).

Schildkamp, Visscher und Luyten (2009) haben fünf Jahre lang 79 niederländische Grundschulen im Rahmen eines freiwilligen Rückmeldeprogramms beobachtet. Innerhalb der fünf Jahre gaben nur jeweils maximal die Hälfte der Schulleitungen und nicht einmal ein Viertel der Lehrkräfte an, dass sie aufgrund der ihnen zugänglichen Daten Maßnahmen zur Qualitätsentwicklung vorgenommen hätten. Eine Wirkung der tatsächlichen Testergebnisse auf die Qualitätsentwicklung konnte nicht nachgewiesen werden (Schildkamp, Visscher & Luyten, 2009). Genauso scheinen bisher Hoffnungen unberechtigt, dass sich durch die Möglichkeiten zur systematischeren Qualitätsentwicklung die Schülerleistungen verbesserten (Visscher & Coe, 2003).

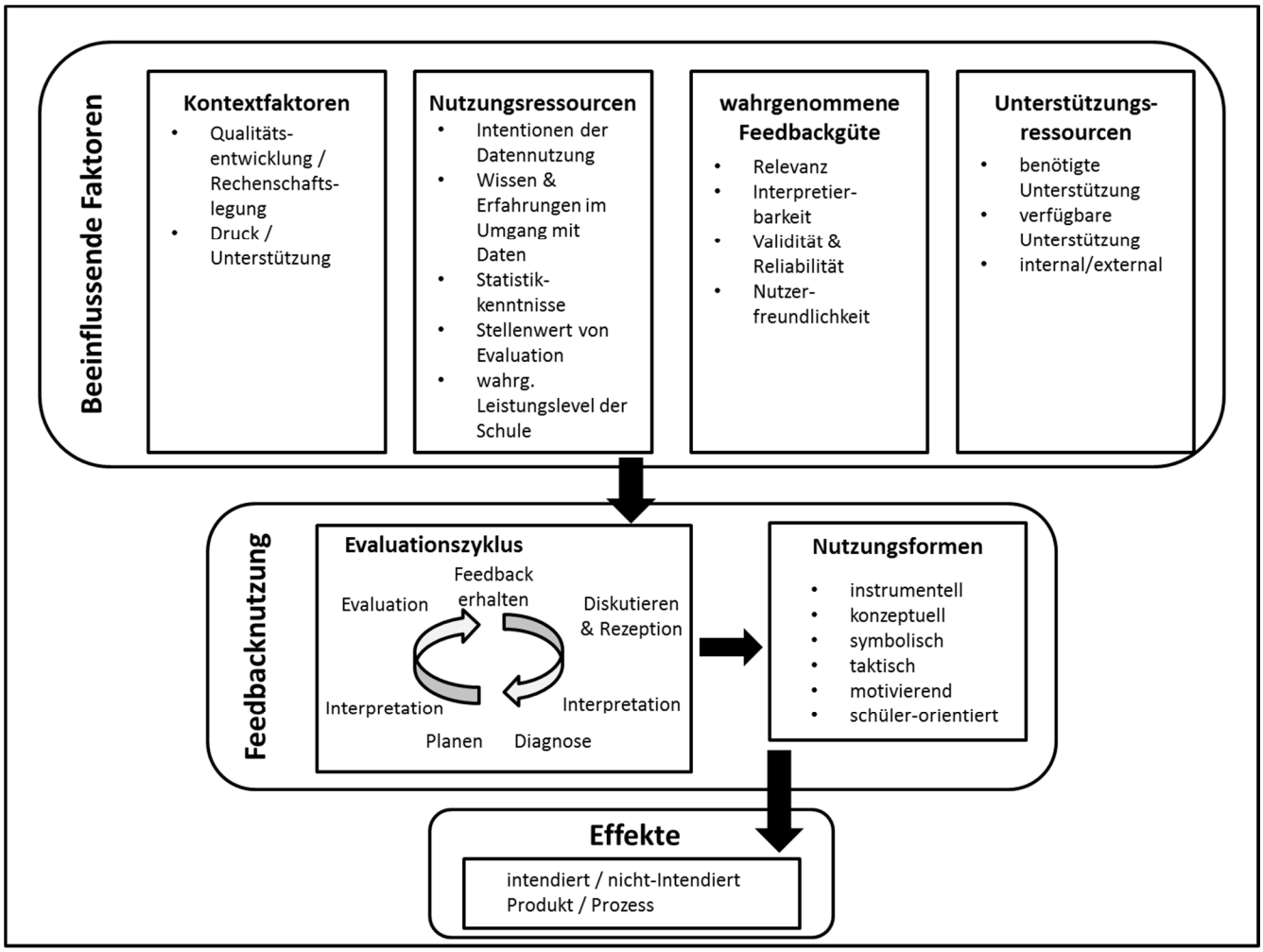


Abbildung 2.1 Modell zu School Performance Feedback Systems nach Verhaegh u.a. (Übersetz. d. Verfass.)

Auch in der Studie von Ingram u.a. (2004) gab nicht einmal die Hälfte der Befragten an, überhaupt Entscheidungen im Kontext von Schuleffektivität auf Grundlage von Ergebnissen aus SSL getroffen zu haben. Überraschenderweise scheint es den deutlichsten Datengebrauch in England zu geben, obwohl dort eher der „Pressure“-Ansatz verfolgt wird. Ashby und Sainsbury untersuchten, ob Testresultate aus SSL in Grundschulen genutzt werden, um das Schulcurriculum zu planen und Schulentwicklung zu betreiben. Sie berichten, dass ungefähr neunzig Prozent der Schulen nach eigenen Angaben Testergebnisse berücksichtigen, um ihr Curriculum aufzustellen, und drei Viertel der Schulen die Resultate auch für die Schulentwicklung nutzen (Ashby & Sainsbury, 2001). Nicht geklärt werden kann aufgrund der Erhebungstechnik, ob dabei tatsächlich eine instrumentelle bzw. konzeptionelle Nutzung vorlag oder ob die Daten nur symbolisch genutzt wurden.

Brown und Harris (2009) berichten aus Untersuchungen von Primarschul- und High School-Lehrkräften mittels Fragebogen und Interviews über den Umgang mit einem Testsystem aus Neuseeland, welches verschiedene Unterstützungsleistungen neben einfachen Tests

anbietet und dem „Support“-Ansatz zuzurechnen ist. Als Ergebnis halten sie fest, dass Lehrkräfte zum Teil das Angebot tatsächlich zur Verbesserung des Lernens heranziehen und entsprechende Maßnahmen berichten. Ein anderer Teil sieht in der SSL hingegen trotz des erkennbaren Unterstützungscharakters vor allem ein Kontrollinstrument. Eine Nutzung findet entweder gar nicht statt oder ist lediglich symbolisch (Brown, 2011; Brown & Harris, 2009).

Über eine sehr hohe Nutzungsquote von Daten aus standardisierten Schulabschlussprüfungen berichtet hingegen Klein (2013) von befragten Schulen in den Niederlanden und etwas geringer durch Schulen in Irland. Sie hat die Nutzung solcher Daten ebenso in Finnland erhoben und kommt zu dem Schluss, dass die Nutzung möglicherweise gerade dort sehr hoch ist, wo durch den Staat ein klarer Nutzungsauftrag formuliert und flankierend zusätzliche Unterstützungssysteme für die Verknüpfung zwischen Prüfungsergebnissen und Unterrichtsqualität existieren (Klein, 2013).

Zusammenfassend zeigen die Erkenntnisse aus der internationalen Forschung ein sehr heterogenes Bild der Wirkung von SSL. Weder der „Pressure“-Ansatz über die betonte Rechenschaftslegung (an dem sich die Steuerung im anglo-amerikanischen Raum eher orientiert) noch der „Support“-Ansatz, der die Qualitätsentwicklung in den Vordergrund stellt, können als Erfolgsmodell gelten. Zwar scheint der „Pressure“-Ansatz bei einem Fokus auf die Verbesserung der Schülerleistung erfolgreicher, gleichzeitig zeigen sich aber auch verschiedene ungewollte Nebenwirkungen. Auch kann aufgrund der Untersuchungsdesigns weder bei der Leistungssteigerung noch bei der Nutzung von Testdaten zur Schul- und Unterrichtsentwicklung ein substantieller Effekt belegt werden. Letzteres kann aber auch der „Support“-Ansatz nicht für sich beanspruchen. Während die hierzu diskutierten Studien Barrieren der datengestützten Schul- und Unterrichtsentwicklung und theoretische Modell aufstellen, bleiben sie den Nachweis schuldig, wann eine entsprechende Datennutzung tatsächlich funktioniert. Offen bleibt auch die Frage, ob solche Erkenntnisse auf überpersonaler Ebene überhaupt möglich sind. Für den „Support“-Ansatz bleibt zusätzlich unbeantwortet, wie nützliche Unterstützungsmaßnahmen aussehen müssen.

Trotzdem wird das bisher Dargestellte helfen, die Anlage von zentralen Vergleichsarbeiten in Deutschland einzuordnen und dadurch Effekte zu antizipieren. Dies soll im restlichen Teil dieses Kapitels geschehen, welcher mit einer knappen Übersicht über Leistungsmessung in Deutschland beginnt.

2.2 Zentrale Lernstandserhebungen in Deutschland

Im zweiten Abschnitt dieses Kapitels werden die Steuerungslogik zentraler Lernstandserhebungen und ihr Zusammenspiel mit den verschiedenen anderen Formen der Schulleistungsmessung in Deutschland behandelt. Zuerst erfolgt daher eine kurze Betrachtung der Schulleistungsmessung in Deutschland insgesamt, anschließend liegt der

Fokus dann auf zentralen Vergleichsarbeiten, der deutschen Variante von zentralen Lernstandserhebungen. Die konkrete Umsetzung von VERA8 in Nordrhein-Westfalen und bisher dokumentierte Steuerungseffekte werden in den Abschnitten (2.3) sowie (2.4) und (4.3.5) thematisiert.

2.2.1 Leistungsmessung im Bundesgebiet

Es sind prinzipiell viele verschiedene Arten der Leistungsmessung in der Schule denkbar.[38] Obwohl nur sehr wenige davon tatsächlich in der Schule genutzt werden (Sacher, 2009), kann eine Kategorisierung nur partiell, beispielsweise für in den Schulen übliche schriftliche Prüfungen, erfolgen.

Die verschiedenen Typen der schriftlichen Leistungsmessung können mit Hilfe von vier Dimensionen unterschieden werden: die Dimension der Aufgabenentwicklung, die Korrektur-, Auswertungs- und Beurteilungs-Dimension, Messbereich-Dimension und die Adressaten-Dimension.

Dimension der Aufgabenentwicklung: Auf welcher Ebene werden die Aufgaben und Erwartungshorizonte entwickelt? Dies kann auf der Klassenebene durch die jeweilige Lehrkraft, auf der Jahrgangsstufenebene in Absprache verschiedener Lehrkräfte, der Ministeriumsebene oder auf überstaatlicher Ebene geschehen. Es können auch mehrere Ebenen beteiligt sein wie bei den Parallelarbeiten (Jahrgangsstufen und Klassenebene) oder dem Zentralabitur (wenn Lehrkräfte für ihre Schüler einige Aufgaben aus einem Aufgabenpool auswählen können). Die Antwort auf diese Frage ist wichtig, weil die Aufgabenentwicklung in die Leistungsattribution einbezogen wird. Weiter ergibt sich heraus, inwieweit die klassischen Testgütekriterien Objektivität, Reliabilität und Validität erfüllt werden können.

Korrektur-, Auswertungs- und Beurteilungs-Dimension: Wer korrigiert die Schülerprodukte und wer nimmt die Auswertung beziehungsweise die Bewertung vor? Der gewöhnliche Fall ist eine Korrektur und Bewertung durch die jeweilige Lehrkraft, aber es sind auch Korrekturen, Auswertungen und Bewertungen auf der Schülerebene, Jahrgangsstufenebene, Schulebene, schulübergreifenden Ebene und auf der Ebene unabhängig-wissenschaftlicher Institutionen wie dem IEA-DPC in Hamburg möglich. Weiter unterscheiden sich die Leistungsmessungen darin, ob Bewertungen nach einer kriterialen, einer sozialen oder einer individuellen Norm vorgenommen werden.

Messbereich-Dimension: Auf welchen Zeitraum erstrecken sich die Aufgaben? Hier reicht die Spannweite von der letzten Unterrichtsstunde bei einem Test über Unterrichtsreihen bei

[38] Sacher rechnet durch Kombination von Leistungsarten, Formen der Inszenierung und Beurteilung vor, dass theoretisch knapp zehn Millionen verschiedene Formen denkbar wären.

Klassenarbeiten, die letzten Schuljahre bis zur kompletten Schulzeit wie bei den Prüfungen am Ende der zehnten Klasse. Auch die Unterscheidung von summativen und formativen Prüfungen fällt unter diese Dimension.

Adressaten-Dimension: An wen richtet sich die Rückmeldung vorwiegend bzw. wer soll ggf. zu Veränderungen angeregt werden? Mögliche Adressaten sind hier Schülerinnen und Schüler, die unterrichtende Lehrkraft, die jeweilige Schule, der Schulträger oder der Staat/das Land.

Bonsen, Büchter und Peek (2006) kategorisieren bei schriftlichen Leistungsmessungen (1) Klassenarbeiten (Stegreiftests/Klausuren), (2) Parallelarbeiten, (3) zentrale Lernstandserhebungen/Vergleichsarbeiten, (4) zentrale (Abschluss-)Prüfungen und (5) Schulsystem-Studien[39]:

Klassenarbeiten (Test/Klausuren) werden vom Fachlehrer entwickelt und ausgewertet. Geprüft wird in der Regel der Inhalt der letzten Unterrichtsreihe, wobei das Ziel die gerechte Bewertung der Fachleistungen der einzelnen Schüler und Schülerinnen ist. Mehrheitlich haben sie den Charakter einer summativen Prüfung, wenngleich dies nicht zwingend ist.
Parallelarbeiten werden vorwiegend jahrgangsstufenweise gemeinsam von den jeweils unterrichtenden Fachlehrern erstellt, teilweise gibt es klassenspezifische Ergänzungen. Die Arbeiten werden von den Fachlehrern selbst oder kreuzweise korrigiert. Prüfungsgegenstand ist der Fachinhalt bis zu des kompletten Schuljahres. Neben der Bewertung der einzelnen Schülerinnen und Schüler dienen sie auch dazu, sich innerhalb einer Jahrgangsstufe über Ziele und Anforderungen auszutauschen.

Zentrale Lernstandserhebungen/Vergleichsarbeiten werden zentral von Fachdidaktikern und Psychometrikern (mehrheitlich koordiniert durch das Institut für Qualitätsentwicklung im Bildungswesen, Humboldt Universität zu Berlin) entwickelt, teilweise wird dabei auf von einzelnen Lehrkräften eingereichte Aufgaben zurückgegriffen. Die Entwickler orientieren sich an curricularen Vorgaben wie beispielsweise den Bildungsstandards. Vorwiegend sollen sie Schul- und Unterrichtsentwicklung anstoßen und sind daher vor allem auf die Möglichkeit hin entworfen, Klassenergebnisse abzubilden. Ergebnisrückmeldungen auf Schulebene werden zusammen mit „Fairen Vergleichen“ angeboten (Bonsen & von der Gathen, 2004). Auch auf Individualebene werden Ergebnisse zurückgemeldet, daher bearbeiten alle Schülerinnen und Schüler landesweit dieselben Aufgaben. Prüfungsgegenstand sind die Unterrichtsinhalte eines Fachs aus mehreren Schuljahren. Die Korrektur der Tests erfolgt schulintern nach einem zentralen Bewertungsschema. Zusätzlich erhofft man sich Systemmonitoring-Wissen durch die jährlich durchgeführten Prüfungen.

[39] Bonsen, Büchter und Peek sprechen von „System-Monitoring-Studien“. Da aber auch zentrale Lernstandserhebungen und zentrale Abschlussprüfungen System-Monitoring-Wissen liefern sollen, ist diese Bezeichnung unglücklich.

Zentrale Abschlussprüfungen werden wie Lernstandserhebungen zentral (durch eine Landesbehörde) entwickelt und prüfen ebenfalls die Unterrichtsinhalte mehrerer Schuljahre. Sie richten sich aber eindeutig an die Schüler und Schülerinnen und dienen der Zertifizierung am Ende der Sekundarstufe I bzw. am Ende der Sekundarstufe II. Das Ziel ist eine Standardisierung der Prüfungen (Kühn, 2010). Gleichzeitig kommt den zentralen Abschlussprüfungen zwar eine bedeutende Funktion bei der Abschlussvergabe zu, vorherige Leistungen werden aber stets mindestens gleichgewichtet berücksichtigt.

Schulsystem-Studien dienen der Bewertung des Schulsystems bzw. von Subsystemen (beispielsweise Schulformen). Die Aufgaben werden zentral auf Landesebene (z.B. LAU oder KESS) oder landes- bzw. staatenübergreifend entwickelt, wenn zusätzlich ein Vergleich mit anderen Bundesländern oder anderen Staaten beabsichtigt ist. Aussagen über einzelne Schülerinnen und Schüler sind in der Regel nicht beabsichtigt, sodass häufig auf ein Multi-Matrix-Design zurückgegriffen wird und nicht alle Testpersonen dieselben Aufgaben bearbeiten. Dadurch kann eine größere Tiefe der Stoffgebiete und eine größere Breite innerhalb der Unterrichtsfächer erfasst werden (Bonsen, Büchter & Peek, 2006; Koch, 2011).

Die tatsächliche Nutzung der fünf schriftlichen Prüfungsformen ist der föderalen Struktur in Deutschland geschuldet sehr unterschiedlich. Während Klassenarbeiten traditionell in allen sechzehn Bundesländern zum Prüfungskanon gehören und sich alle Bundesländer vor allem in den letzten 15 Jahren an Schulsystem-Studien beteiligten, existiert für zentrale Lernstandserhebungen, Parallelarbeiten und zentrale Abschlussprüfungen bisher kein alle Länder umfassender Konsens. Nachdem mittlerweile nach den Ländern der ehemaligen DDR, Bayern, Baden-Württemberg und dem Saarland auch fast alle anderen Länder zentrale Prüfungen am Ende der Sekundarstufen I und II eingeführt haben, bleibt formal nur Rheinland-Pfalz außen vor. In der konkreten Umsetzung bzgl. der Anzahl und Auswahl der Prüfungsfächer und der Korrektur gibt es allerdings auch innerhalb dieser 15 Länder deutliche Differenzen (van Ackeren, 2007; Klein, Kühn, & Block, 2009). Parallelarbeiten existieren ebenfalls nicht in allen Bundesländern. Zentrale Lernstandserhebungen werden zwar in allen sechzehn Bundesländern geschrieben, die Teilnahme ist aber nicht in jedem Bundesland in allen Fächern verbindlich. Einige Bundesländer bieten stattdessen zentrale Klassenarbeiten an, die wie Parallelarbeiten genutzt werden können. Außerdem variiert die Zahl der zentralen Lernstandserhebungen unter den Bundesländern zwischen zwei (Berlin, Brandenburg, Bremen, Niedersachsen, Nordrhein-Westfalen, Rheinland-Pfalz, Sachsen-Anhalt, Saarland, Schleswig-Holstein) und drei (Baden-Württemberg inkl. Diagnosearbeiten in Jgst. 7+9, Bayern inkl. Orientierungsarbeit in Jgst. 2, Hamburg, Hessen, Mecklenburg-Vorpommern, Sachsen, Thüringen). Nur in Sachsen werden neben den Fächern Deutsch, Mathematik und Fremdsprachen auch die Naturwissenschaften durch zentrale Lernstandserhebungen evaluiert, Gesellschaftswissenschaftliche Fächer werden in keinem Bundesland berücksichtigt (vgl. Tab. 2.1).

Tabelle 2.1

Übersicht über zentrale Lernstandserhebungen in Deutschland

Bundesland	Jahrgangsstufe							
	2	3	4	5	6	7	8	9
Baden-Württemberg	-	D, M VERA3 vT	-	-	-	D,M DVA	-	D,M, 1.FS (nicht HS und WRS) DVA
Bayern	R OA vT	D, M VERA3 vT	-	-	D, M, E, L (*nur GY*) zK vT	-	D, M, E VERA8 vT	-
Berlin & Brandenburg	-	D, M VERA3 vT	-	-	-	-	D, M, 1.Fs VERA 8 vT in einem Fach, wechselnd	-
Bremen	-	D, M VERA3 vT	-	-	-	-	D, M, E VERA 8 vT in einem Fach, wechselnd	-
Hamburg	-	D, M VERA3 vT	-	-		-	D, M, 1.Fs VERA 8 vT	-
Hessen	-	D, M VERA3 vT	-	-	D, M, E Lernstand6 fw	-	D, M, 1.Fs VERA 8 vT in einem Fach	-
Mecklenburg-Vorpommern	-	D, M VERA3 vT	-	-	D, M, E Lernstand6 vT	-	D, M, E VERA 8 vT	-
Niedersachsen	-	D, M VERA3 in Teilen verpflichtend	-	-	-	-	D, M, 1. Fs VERA 8 in Teilen verpflichtend	-

Bundesland	Jahrgangsstufe							
	2	3	4	5	6	7	8	9
Nordrhein-Westfalen	-	D, M mit VERA3 vT	-	-	-	-	D, M, 1.Fs mit VERA 8 vT	-
Rheinland-Pfalz	-	D, M mit VERA3 vT	-	-	-	-	D, M, E VERA 8 vT in einem Fach, wechselnd	-
Saarland	-	D, M VERA3 vT	-	-	-	-	ein Fach VERA8 vT	-
Sachsen	-	D, M VERA3 vT	-	-	D, M, E Lernstand6 vT	-	D, M, E VERA8 vT	-
Sachsen-Anhalt	-	D, M mit VERA3 vT	D,M zK vT in einem Fach		D,M, E zK vT in einem Fach	-	wechselnd VERA8 + NW vT	D,M
Schleswig-Holstein	-	D, M VERA3 vT	-	-	-	-	D,M, E VERA8 vT	
Thüringen	-	D, M VERA3 vT	-	-	D, M, E Lernstand 6 vT in einem Fach	-	D,M, 1.Fs VERA8 vT	-

Legende	**Fächer:** Deutsch (D), Mathematik (M), Englisch (E), 1. Fremdsprache (1.Fs), Rechtschreib. (R), Latein (L), Naturwissenschaften (NW) **Schulform:** Gymnasium (GY) Hauptschule (HS), Realschule (RS), Werkrealschule (WRS) **Instrument:** Diagnosevergleichsarbeit (DVA), Orientierungsarbeit (OA) zentrale Klassenarbeit (zK), Diagnosetest (DT) **Teilnahme:** verpflichtende Teilnahme (vT), freiwillige Teilnahme (fw)
Quelle	eigene Recherche sowie persönliche Informationen der zuständigen Stellen aus Niedersachsen und dem Saarland – Stand: Juli 2013

Die Tabelle verdeutlich, dass die Bundesländer den im Jahre 2006 von der Kultusministerkonferenz getroffenen Beschluss, jährlich flächendeckende, an die Bildungsstandards angelehnte Lernstandserhebungen in der Sekundarstufe I durchzuführen (Köller, 2008), verschieden umsetzen. Die Lernstandserhebungen unterscheiden sich bezüglich des Zeitpunkts in der Schullaufbahn und der betroffenen Fächer. Legt man darüber hinaus die beiden von Schrader und Helmke genannten Arten von Beurteilungsleistungen zugrunde – (1) messen und beurteilen von Schülerleistungen und (2) messen und beurteilen des Unterrichtserfolgs (Schrader & Helmke, 2002) -, stellt man außerdem eine Schwerpunktverschiebung zwischen den einzelnen Bundesländern fest (Hovestadt & Kessler, 2005). In einigen Bundesländern dürfen die Ergebnisse aus Vergleichsarbeiten auch zur Notengebung herangezogen werden (z.B. VERA8 in Nordrhein-Westfalen), in anderen ist dies hingegen nicht vorgesehen (z.B. VERA8 in Bayern, Berlin und Brandenburg). Dies kann als Zeichen dafür gesehen werden, dass man sich noch nicht sicher ist, welchen Platz zentrale Lernstandserhebungen im Zusammenspiel mit den vier anderen schriftlichen Prüfungsformen einnehmen. Trotzdem gibt es eine starke Tendenz, zentrale Lernstandserhebungen als eigenständiges Instrument zu etablieren. Dieser Weg unterscheidet sich von Qualitätssicherungsverfahren mittels High-Stake-Tests wie sie in (2.1.2) beschrieben wurden und gleicht eher dem in (2.1.3) dargestellten Vorgehen. Dieser Schluss gilt insbesondere für die nachfolgend dargestellten Vergleichsarbeiten in der achten Klasse (VERA8).

2.2.2 Zentrale Vergleichsarbeiten in Klasse 8 (VERA8) in Deutschland

Die zentralen Vergleichsarbeiten in Klasse 8 (VERA8) werden seit dem Schuljahr 2008/2009 von fünfzehn teilnehmenden Bundesländern (Baden-Württemberg nimmt nicht teil) durchgeführt. Vorher existierten unterschiedliche Vorläuferprojekte, bei denen nur einzelne Länder kooperierten (z.B. bei Lernstand8 Mecklenburg-Vorpommern mit Nordrhein-Westfalen). Die Länder sind für die Durchführung zuständig und organisieren die Vorbereitung (z.B. die Verteilung der Testhefte), den Ablauf, die Auswertung und die Ergebnisrückmeldung individuell. Teilweise existieren dazu eigene Stellen in den jeweiligen Ministerien, teilweise gibt es eine Kooperation mit der Universität Koblenz-Landau für die Auswertung und Ergebnisrückmeldung. Der Erhebungszeitpunkt liegt für alle teilnehmenden Länder im Zeitraum Ende Februar bis Anfang März. Die Tests werden aber nicht bundesweit in allen Bundesländern am selben Tag durchgeführt. Der Name „Vergleichsarbeiten“ ergibt sich aus dem Beurteilungsmaß der Ergebnisse: Statt an einer Sachnorm orientiert sich die Rückmeldung in den einzelnen Ländern vor allem an einem Vergleich mit dem Landesdurchschnitt und mit anderen Klassen der achten Jahrgangsstufe der jeweiligen Schule (soziale Bezugsnorm). Dabei wird ein so genannter „Fairer Vergleich“ angeboten, bei dem Kontextmerkmale wie das soziales Umfeld der Schule oder der Anteil an Schülerinnen und Schülern mit Migrationshintergrund berücksichtigt werden.

Die Aufgabenentwicklung[40]

Die Aufgabenentwicklung wird vom Institut für Qualitätsentwicklung im Bildungswesen (IQB) der Humboldt Universität zu Berlin koordiniert. Mit der Entwicklung der Aufgaben sind erfahrene Lehrkräfte betraut. Anschließend werden die Aufgaben von Fachdidaktikern und Testspezialisten des IQB überprüft und einem Pretest unterzogen. Es werden Test für die Fächer Deutsch, Mathematik, Englisch und Französisch (die letzten beiden als erste Fremdsprache) entwickelt. Die vorgesehene Bearbeitungszeit beträgt in jedem Fach 80 Min. Es sind Testhefte in drei Schwierigkeitsniveaus verfügbar, wobei die Bundesländer jeweils entscheiden, an welcher Schulform welches Schwierigkeitsniveau zugrunde gelegt wird. Dadurch soll das unterschiedliche Leistungsniveau der Schulformen genauer erfasst werden. Die Testhefte enthalten aber auch gemeinsame Aufgaben. Die verwendeten Aufgaben zeichnen sich dadurch aus, dass sie als „Aufgaben zum Leisten“ konzipiert sind, um so gezielt bestimmte Kompetenzen erheben zu können. Die Aufgaben unterscheiden sich dadurch von „Aufgaben zum Lernen“, die sich kreativer bearbeiten lassen und offener gestaltet sind, und entsprechen damit nicht dem Aufgabentyp, welcher in den Kernlehrplänen für den Unterricht favorisiert wird (Büchter & Leuders, 2005). Der in den Lernstandserhebungen vorkommende Aufgabentyp sollte den Schülerinnen und Schülern folglich wesentlich seltener begegnet sein und ihnen vorwiegend aus den Phasen des Unterrichts bekannt sein, in denen es um Lernerfolgskontrolle geht. Um in der geringen Testzeit möglichst differenzierte Kompetenzen erfassen zu können, werden in den Fächern Deutsch, Englisch und Französisch nicht in jedem Jahr dieselben Teilkompetenzen (z.B. Hörverstehen) getestet, sondern es findet ein Wechsel von Jahr zu Jahr statt. Die jeweiligen Bereiche werden einige Wochen vor dem Testdatum bekannt gegeben. Dadurch ist es Lehrkräften möglich, ggf. noch Anpassungen ihres Klassencurriculums vorzunehmen. Andersherum ermöglicht dies aber auch eine thematische Einschränkung. Aus diesem Grund werden bisher in Mathematik alle Bereiche getestet.

Da die Aufgaben von Lehrkräften ausgewertet werden, ist es möglich, Aufgabenformate einzusetzen, die sich nicht nur auf Antwortformate wie Multiple-Choice-Aufgaben oder einzelne Zahlen und Wörter beschränken, sondern es können auch offene Antwortformate eingesetzt werden. Trotzdem müssen auch diese Aufgaben eine Differenzierung nach ihrem Schwierigkeitsgrad zulassen und einfach auswertbar sein (Büchter & Leuders, 2005; Burkard & Peek, 2004). Welche Aufgabentypen in einem Jahr Bestandteil der Lernstandserhebungen sind, ist bis zur Auslieferung der Testhefte unbekannt. Gewisse Rückschlüsse lassen aber die Aufgaben in den früheren Testheften und die Beispielaufgaben auf der Internetseite des IQB zu.

Für das Fach Mathematik seien die Aufgaben hier kurz beschrieben: Bei den Aufgaben werden Kontextinformationen durch einen Text, eine Zeichnung ein Foto oder durch eine Tabelle gegeben. Lösungsvarianten sind eine Zahl, mehrere Zahlen, eine zeichnerische

[40] Die Informationen zu Aufgabenkonstruktion, Durchführung und zur Verfügung gestellten Materialien sind der Internetseite des IQB entnommen, so weit nicht anders gekennzeichnet.

Begründung, eine rechnerische Begründung oder ein Text als Begründung, das Ergänzen von Zeichnungen bzw. Tabellen oder die Auswahl aus Antwortalternativen (Multiple-Choice-Aufgaben). Die meisten Aufgabenformate sind aus den aktuellen Schulbüchern vertraut. Jedoch können Multiple-Choice-Aufgaben, aber auch Begründungen schriftlich zu notieren statt sie nur verbal zu geben, könnten für Schüler und Schülerinnen unbekannt sein. Inhaltlich werden mit den Aufgaben die fünf Leitideen der Bildungsstandards (Zahl, Messen, Raum und Form, Funktionaler Zusammenhang, Daten und Zufall - Blum, 2010) abgedeckt.

Neben den Aufgaben erstellen die Entwickler auch „didaktische Handreichungen". Diese enthalten Hinweise und Kommentare zu den eingesetzten Aufgaben und werden auf der Homepage des IBQ und der entsprechenden Internetseiten der Länder nach der Durchführung zur Verfügung gestellt, um die unterrichtenden Lehrkräften im Bemühen um einen kompetenzorientierten Unterricht zu unterstützen. Die Handreichungen sollen außerdem die Nutzungsmöglichkeiten der Testaufgaben und die Grenzen der Vergleichsarbeiten im kompetenzorientierten Unterricht verdeutlichen sowie förderdiagnostische Hinweise geben. Insgesamt dienen die Handreichungen dazu, die Professionalisierung von Lehrkräften zu erweitern und Schul- bzw. Unterrichtsentwicklungsprozesse zu unterstützen.

Die Bildungsstandards

Die inhaltliche Basis für die zentralen Vergleichsarbeiten sind die Bildungsstandards. Bildungsstandards bilden in Deutschland die Zielorientierung der Neuen Steuerung. Sie stellen die Input-Komponente dar und sollen die notwendige Orientierung geben. Gleichzeitig sollen sie Lernergebnisse messbar machen, um Systemmonitoring und Schulevaluation zu ermöglichen bzw. zu unterstützen (Klieme et al., 2003). Auch verspricht man sich von der Einführung von Bildungsstandards eine gerechtere Leistungsbewertung, da Bildungsstandards prinzipiell eine Bewertung anhand einer Sachnorm ermöglichen und dadurch die üblicherweise von Lehrkräften herangezogene Sozialnorm ablösen können (Hartung-Beck, 2009; Schrader & Helmke, 2002). [41] Böttcher (2006) stellt vier Qualitätskriterien für Bildungsstandards auf: Klarheit, Knappheit, Realismus und Anspruch. Klarheit meint, dass die Standards ausreichend präzise und für Lehrkräfte, Erziehungsberechtigte und Schülerinnen und Schüler verständlich sind, um abschätzen zu können, was gelernt werden soll. Sind Standards interpretationsbedürftig, hält Böttcher eine Variation auf Unterrichtsebene für zwangsläufig, sodass (eine Überprüfung und damit überhaupt) Steuerung möglich wird. Gleichzeitig müssen die Standards so weit begrenzt werden (Knappheit), dass keine Auswahl getroffen werden muss und dadurch eine Varianz entsteht. Realismus und Anspruch stecken den inhaltlichen Qualitätsrahmen ab. Die

[41] Diese Hoffnung ist allerdings mit der Formulierung von Regelstandards und der Nutzung von zentralen Lernstandserhebungen als Vergleichsarbeiten schwierig vereinbar.

Standards dürfen weder unerreichbar noch zu einfach erfüllbar sein. Böttcher nennt Standards, die diese Kriterien erfüllen, „starke Standards“ (Böttcher, 2006).

Mit den Bildungsstandards sollen das schulfachliche Lernen und allgemeine Bildungsziele verbunden werden (Maag Merki, 2010). Sie verknüpfen in der Theorie den mit zentralen Tests (zentralen Vergleichsarbeiten und zentralen Abschlussprüfungen in Deutschland) messbaren Output mit dem intendierten Outcome, indem sie statt fachlicher Inhalte Kompetenzen definieren. Die Bildungsstandards zielen somit auf kumulatives und systematisch vernetztes Wissen und definieren grundlegende Kompetenzen in den einzelnen Fächern, die Schülerinnen und Schüler zu vier Zeitpunkten beherrschen sollen (Klieme et al., 2003): zum Ende der Primarstufe[42], zum Zeitpunkt des Hauptschulabschlusses, zum Zeitpunkt des Mittleren Schulabschlusses und zum Zeitpunkt der Allgemeinen Hochschulreife. Bildungsstandards existieren für die Fächer Deutsch, Mathematik (für alle vier Zeitpunkte), Englisch bzw. Französisch als erste Fremdsprache (Hauptschulabschluss und Mittlerer Schulabschluss) bzw. als fortgeführte Fremdsprache (Allgemeine Hochschulreife) und für die drei Naturwissenschaften (nur Mittlerer Schulabschluss). Für die Darstellung werden die Anforderungen in Kompetenzmodellen systematisch in Kompetenzstufen angeordnet, die Abstufungen und Entwicklungsverläufe verdeutlichen, und in Aufgaben und Testverfahren illustriert (Maag Merki, 2010). Mit der Einführung von Bildungsstandards verfolgen die sechzehn Bundesländer eine gemeinsame Strategie, bei der durch Bildungsstandards die Qualität schulischer Bildung, die Vergleichbarkeit der schulischen Abschlüsse und die Durchlässigkeit des Bildungssystems gesichert werden sollen (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2005).

Sowohl Wissenschaftler als auch Lehrervertreter haben die im Jahr 2004 verabschiedeten Bildungsstandards kritisiert. Mit Rückgriff auf Huber, Späni, Schmellentin und Criblez (2006) lässt sich die Kritik in diese Punkte gliedern: (1) fehlende Qualitätssicherung und -entwicklung: Ursprünglich sollten Mindeststandards formuliert werden (Huber, Späni, Schmellentin & Criblez, 2006), verabschiedet wurden aber im Jahr lediglich Regelstandards, da keine empirische Validierung vorlag (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2005). Befürchtet wird daher, dass die Standards als durchschnittliche Ziele nicht nur einige Schüler und Schülerinnen über- und andere unterfordern, sondern dass das Ziel nicht erreicht wird, die Qualität der schulischen Bildung zu sichern. Regelstandards sind weder verbindlich noch können sie entsprechend überprüft werden. Durch die Definition für das Ende eines Schulabschnitts wird auch der Entwicklungsgedanke nicht ausreichend illustriert. (2) Fehlende Orientierung: Die Zahl der Standards ist für die einzelnen Fächer sehr hoch, sodass gerade keine Konzentration auf die wesentlichen Inhalte der Fächer vorliegt. Sie unterscheiden sich somit nur gering von den früheren Lehrplänen und sind schlecht operationalisiert. Im Sinne von Böttcher können sie folglich nicht als „starke Standards“ bezeichnet werden. (3) Falsche

[42] Es wird eine vierjährige Primarstufe angenommen.

Akzentuierung: Die Standards sind nur für so genannte Kernfächer definiert worden. Dies könnte zu Vernachlässigung von „weichen Fächern“ führen. Außerdem sind keine fachübergreifenden Kompetenzen erfasst worden (Huber et al., 2006). Folgt man Böttcher, muss in Frage gestellt werden, ob die Bildungsstandards eine ausreichende Basis für Selbstevaluationsprozesse an Schulen darstellen. Sind Standards nicht stark genug, sind nur Verfahren externer Evaluation möglich (Böttcher, 2006).

Die Intentionen bei VERA8

Mit der Einführung von zentralen Lernstandserhebungen in Deutschland sind grob vier Ziele verbunden: die Schul- und Unterrichtsentwicklung als wichtigstes Ziel, die Standardsicherung und -entwicklung (auch im Sinne einer Bestandsaufnahme der Kompetenzverteilung in Deutschland), die weitere Professionalisierung von Lehrkräften sowie Informationen für die Erziehungsberechtigten (Koch, 2011). Zentrale Vergleichsarbeiten fügen sich dadurch sehr gut in das in (2.1.1) beschriebene Steuerungskonzept, indem die SSL eine wichtige Rolle erhält. Von den im Abschnitt (2.1.1) beschriebenen acht steuerungstheoretischen Funktionen wird nur die achte Funktion, Wettbewerb zu ermöglichen, nicht aufgegriffen. Alle anderen sieben Funktionen (Kühn, 2010; Maritzen, 2008; Tresch, 2007) der Instrumente zur Gewinnung von Daten aus externen Evaluationen lassen sich unter den von Koch genannten Zielen verorten.

Zur gleichen Zeit sollen mit SSL Aufgaben der Bildungsadministration umgesetzt und Lehrkräften und Schulen eine Weiterentwicklung ermöglicht werden. Die von Burkard und Peek (2004), Helmke und Hosenfeld (2003), Heymann und Pallack (2007) sowie Parveva et al. (2009) formulierten Funktionen beziehen sich explizit auf die zentralen Vergleichsarbeiten aus dem Projekt VERA8 und VERA3. Für die prognostische Funktion kann das allerdings nicht für alle Bundesländer geltend gemacht. Anders als zentrale Abschlussprüfungen und Qualitätsprüfungsinstrumente in anderen Staaten weisen zentrale Vergleichsarbeiten in ihrer Grundidee keine Allokations-, Disziplinierungs-, Selektions- oder Sozialisationsfunktion auf, wenngleich dieses in einigen Bundesländern dadurch aufgeweicht wurde, dass die Ergebnisse in Ausnahmefällen in die Zeugnisnote einfließen können oder die Ergebnisse aus VERA3 bei der Schullaufbahnempfehlung berücksichtigt werden (Maag Merki, 2010).

Offen bleibt dabei allerdings die Gewichtung der verschiedenen Funktionen bzw. der verschiedenen Ziele. Koch (2011) merkt an, dass es dabei über die Jahre zu Verschiebungen gekommen ist. Die bereits angesprochene Funktionsüberfrachtung wird gerade bei VERA8 deutlich. VERA8 soll als Instrument gleichzeitig neue Inhalte und Aufgabenformate implementieren und die Diagnose von Lehrkräften unterstützen. Eine richtige Diagnose setzt aber auch eine richtige Attribution der Schülerleistung voraus. Nur wenn die neuen Inhalte und vor allem die neuen Aufgabenformate bereits Teil des regulären Unterrichts sind, werden Abweichungen zur erwartbaren Leistung und Unterschiede zwischen einzelnen Schülerinnen und Schülern wahrgenommen und angemessen reflektiert. Genauso existiert

zweitens möglicherweise eine Aporie durch die deskriptive und die innovative Funktion. Wenn die Lernstandserhebungen die Erfolge des bisherigen Unterrichts beschreiben sollen, müssen die in den Lernstandserhebungen verwendeten Aufgaben mit dem Unterricht kohärent sein. Ist der Unterricht mit den Aufgaben der Lernstandserhebungen aber schon kohärent, können die Aufgaben nicht mehr dazu anregen, Unterrichtsentwicklung in diese Richtung zu betreiben. Schließlich kann auch noch zwischen den dem Projekt VERA8 von Wissenschaftlern, Politikern und Lehrkräften unterschiedlich zugewiesenen Funktionen differenziert werden. Während beispielsweise die Erstellung von Schulrankings immer wieder von Politikern und Wissenschaftlern bestritten und ihre Realisierbarkeit mittels zentraler Vergleichsarbeiten wie VERA8 angezweifelt wird, hält ein Teil der Lehrkräfte diese Nutzung für gegeben (Hahn, 2008). Vor allem die Einschätzung der Lehrkräfte von VERA8 kann als wichtige Komponente bei Fragen nach einer Vorbereitung auf VERA8 betrachtet werden. Die transportierten Intentionen als Gesamtbild hängen auch von den Kommunikationswegen, der Ergebnisrückmeldung und von dem Umgang mit den Ergebnissen für die Individualbewertung der Schülerleistungen ab. Da dies in die Verantwortung der Bundesländer fällt und die beiden Studien dieser Arbeit in Nordrhein-Westfalen stattfanden, werden diese nachfolgend für Nordrhein-Westfalen dargestellt.

2.2.3 VERA8 in Nordrhein-Westfalen[43]

Zentrale Lernstandserhebungen werden in Nordrhein-Westfalen seit 2004 durchgeführt und firmieren unter dem Projektnamen „Lernstand8“ bzw. vorher „Lernstand9“. Lehrkräfte in Nordrhein-Westfalen gehören somit zu den Lehrkräften der Sekundarstufe I mit der größten Erfahrung im Umgang mit zentralen Lernstandserhebungen. In den Schuljahren 2004/05 und 2005/06 nahmen jeweils im November alle neunten Klassen der Haupt-, Real- und Gesamtschulen sowie Gymnasien in Nordrhein-Westfalen an den Tests in Deutsch, Englisch und Mathematik teil. Anschließend wurde der Erhebungszeitpunkt modifiziert und in den Schuljahren 2006/07 und 2007/08 wurden die Lernstandserhebungen in den achten Klassen der entsprechenden Schulformen zu Beginn des Sommerhalbjahres durchgeführt.[44] Seit dem Schuljahr 2008/09 beteiligt sich Nordrhein-Westfalen an VERA8. Neben Englisch können Schulen, die Französisch ab der fünften Klasse unterrichten, auch Vergleichsarbeiten für Französisch nutzen.

[43] Die nachfolgenden Ausführungen beziehen sich auf die Umsetzung von VERA8 als Lernstand8 in Nordrhein-Westfalen für das Jahr 2010 bzw. auf die Jahre davor. Modifikationen wurden mittlerweile beispielsweise bei der Ergebnisrückmeldung und der Konzeption der „Fairen Vergleiche“ („Standorttypen“) vorgenommen.

[44] Durch den modifizierten, früheren Erhebungszeitpunkt sollte eine bessere Möglichkeit gegeben werden, in den Lernstandserhebungen entdeckte Defizite der Schüler zu beheben. Die Tests waren von Beginn bis 2008 für die Jahrgangsstufen 5 bis 8 konzipiert, sodass die Tests ursprünglich nach dem Zeitpunkt durchgeführt wurden, zu dem die Schüler die abgefragten Kompetenzen bereits besitzen sollten. Das Auswertungsprozedere sieht eine zeitnahe Ergebnisrückmeldung auf Schul- und Klassenebene vor. Auch wenn die vollständige Auswertung erst jeweils nach den Sommerferien abgeschlossen ist, bleibt den Lehrkräften noch einige Zeit, ggf. vorhandene Defizite zu korrigieren.

Information und Kommunikation

Bei der Kommunikation der Projektverantwortlichen mit den Schulen und der restlichen interessierten Öffentlichkeit spielt das Internet eine wichtige Rolle. Über die Internetseite werden allgemeine Informationen zu Zielen, Abläufen und Ergebnissen gegeben und häufig gestellte Fragen beantwortet. Daneben wurden bis 2008 Beispielaufgaben mit didaktischen Handreichungen angeboten. [45] Außerdem werden LSE-Koordinatoren (LSE für Lernstandserhebung) einer Schule per E-mail über wichtige Neuigkeiten zu den Lernstandserhebungen informiert und die Datenübermittlung online abgewickelt. Um möglichen Problemen aus dem Weg zu gehen, wurde außerdem eine Hotline eingerichtet, an die sich Lehrerinnen und Lehrer mit Fragen zu den Lernstandserhebungen richten können.

Die Erziehungsberechtigten der betroffenen Schülerinnen und Schüler werden zusätzlich in einem an ihre Kinder verteilten Anschreiben darüber informiert, welche Ziele mit den Lernstandserhebungen verfolgt werden, wie die Lernstandserhebungen ablaufen und welche Konsequenzen das Abschneiden des Kindes haben kann. Darüber hinaus werden die Erziehungsberechtigte gebeten, ihr Kind zu bestärken, die Lernstandserhebungen nach bestem Vermögen zu absolvieren. Für weitere Informationen wird auf die Internetseite verwiesen (Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen). Das Anschreiben hat den Umfang von zwei Seiten und ist in Deutsch, Bosnisch, Kroatisch, Serbisch, Polnisch, Russisch und Türkisch verfügbar.

Über das jeweilige Abschneiden der Schulen berichten nach Veröffentlichung der Ergebnisse durch die Schulen die regionalen Zeitungen. Ob sie Ergebnisse veröffentlichen, entscheiden die jeweiligen Schulen individuell. Die besten Schulen - unterschieden nach Schulform, Standorttyp und Fach - wurden von 2007 bis 2009 von der Ministerin ausgezeichnet, welches medienwirksam inszeniert ebenfalls Bestandteil der Berichterstattung von Tageszeitungen, Magazinen und Fernsehsendern war. Erziehungsberechtigte erhalten daher neben der Rückmeldung durch die Schule (s.u.) auch Informationen über die Ergebnisse anderer Schulen aus der Region und könnten somit zu einem anderen Bild der Schule ihrer Kinder gelangen, als die schulische Administration für Lernstand8 beabsichtigt.

Rückmeldungen

Die Rückmeldung bei Lernstand8 erfolgt jährlich in zwei Phasen. Direkt nachdem die Ergebnisse der Schüler und Schülerinnen durch die jeweiligen Lehrkräfte eingegeben worden sind, ist online eine Ergebnisrückmeldung abrufbar, die einen Überblick über die jeweilige Klasse bzw. den Kurs bietet. Dies ermöglicht dann die Leistungen der Schülerinnen und Schüler zu reflektieren, Fördermaßnahmen einzuleiten und im Sinne des Erlasses (s.u.) bei der Vergabe der Zeugnisnoten zu berücksichtigen. Zu diesem Zeitpunkt ist es noch nicht

[45] Seit dem Schuljahr 2008/09 liegt dies in der Verantwortung des IQB.

möglich, die Klassenergebnisse im Vergleich zu Parallelklassen oder gar zu anderen Schulen in Beziehung zu setzen, da deren Ergebnisse evtl. noch nicht eingetragen wurden. Die Ergebnisse der ganzen Klasse können also nur mit einer kriterialen Norm eingeschätzt werden. In der Regel werden die Testhefte den Schülerinnen und Schülern in einer Unterrichtsstunde pro Fach zur Ansicht überlassen und die Aufgaben durch die Fachlehrkräfte besprochen. In diesem qualitativen Teil können die didaktischen Handreichungen genutzt werden, in denen Hinweise auf typische Fehler und Vorstellungen von Schülerinnen und Schüler gegeben werden und weiteres Arbeiten mit ähnlichen Aufgaben vorgeschlagen wird (Büchter & Leuders, 2005).

Diese erste Phase entspricht im Prinzip der Phase, die im Anschluss an eine Klassenarbeit folgt. Erst einmal stehen hier die individuellen Leistungen der Schüler und Schülerinnen im Mittelpunkt. Diese können natürlich sehr wohl auch mit einer Sozialnorm bewertet werden, denn die Klassenergebnisse liegen als Referenzrahmen vor. Für die Lehrkräfte können die Ergebnisse zusammen mit der angebotenen Kommentierung der Aufgaben einen Anreiz bieten, den eigenen Unterricht zu überdenken, sie müssen aber zu diesem Zeitpunkt darüber keine Rechenschaft ablegen.

Die zweite Phase der Rückmeldung erfolgt erst im neuen Schuljahr. Diese zweite Rückmeldung bietet im quantitativen Teil eine Übersicht über eine Einordnung in Kompetenzniveaus und Vergleiche zu anderen Schulen gleichen Schultyps (Büchter & Leuders, 2005). Dabei wird nicht nur zwischen den Schulformen (und bei Gesamtschulen zwischen Erweiterungs- und Grundkurs) unterschieden, sondern auch zwischen Standorttypen (zwei verschiedene Typen bei Realschulen und Gymnasien – drei bei Haupt- und Gesamtschulen). Die Schulen haben sich selbstständig vor den Erhebungen einem Standorttyp zugeordnet, indem sie angegeben haben, ob bestimmte Kriterien (zu Migrationshintergrund/Deutschkenntnissen, Einkommenssituation der Erziehungsberechtigte und des Wohnumfelds, Bildungshintergrund der Erziehungsberechtigte, bei HS zus. Gesamtschule im Einzugsgebiet) auf ihre Schülerinnen und Schüler zutreffen. Mittlerweile werden schulformübergreifend fünf Standorttypen unterschieden. Die Zuordnung erfolgt dabei nicht mehr durch die Schule selbst, sondern wird aufgrund von sozioökonomischen Daten durch das MSW vorgenommen.

Auch wenn in dieser zweiten Phase eine Rückmeldung an die Erziehungsberechtigten über die individuellen Leistungen ihrer Kinder, die Leistungen der Klasse und das Abschneiden der Schule erfolgt, steht nun mehr die Qualitätskontrolle des Unterrichts im Mittelpunkt. Die Fachkonferenzen sind angewiesen, die Ergebnisse zu reflektieren und mögliche Konsequenzen für den Unterricht zu erarbeiten. Die Reflexion und die geplanten Konsequenzen müssen bis zum Ende des ersten Schulhalbjahres an die Schulaufsicht berichtet werden. Um diese Arbeit zu erleichtern und um einen Mindeststandard der Reflexion abzusichern, gibt es dazu Anleitungen und ein standardisiertes Auswertungs- und Berichtsraster (Orth, 2005). Zusätzlich informiert die Schulleitung (der LSE-Koordinator/die LSE-Koordinatorin) über die Ergebnisse der Schule in der Schulkonferenz. Erst nach dem

ersten Schulhalbjahr werden die Testhefte endgültig wieder an die Schülerinnen und Schüler zurückgegeben.

Die Bewertung der individuellen Leistungen

VERA8 wird an Stelle einer früher vorgesehenen Klassenarbeit geschrieben. Um den Aspekt der Qualitätssicherung und -entwicklung herauszustellen, sollen die Leistungen der Schülerinnen und Schüler aber nicht benotet werden. Seit 2007 sollen die Lehrkräfte die Ergebnisse der Lernstandserhebungen daraufhin beurteilen, ob die Schülerinnen und Schüler im Vergleich zu den bisherigen Leistungen bessere, gleiche oder schlechtere Leistungen abgeliefert haben, und diese Beurteilungen sollen bei der Entscheidung über die Zeugnisnoten berücksichtigt werden, wenn zwischen zwei Noten entschieden werden muss (Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, 2007). Der Stellenwert der Lernstandserhebungen für die individuelle Schullaufbahn eines Schülers oder einer Schülerin ist durch die Regelung möglicherweise weiterhin nicht dergleiche wie der Stellenwert einer Klassenarbeit. Der ursprünglich intendierte Effekt, durch einen Verzicht auf eine Bewertung der Schülerleistung als Rückmeldung an die Schülerinnen und Schüler die Unterrichtsentwicklung mit den Lernstandserhebungen zu fokussieren, wird durch diesen Beschluss allerdings untergraben.

2.2.4 Ein Modell zur Nutzung von Ergebnisrückmeldungen aus zentralen Vergleichsarbeiten im deutschsprachigen Raum

Studien zur Nutzung von Ergebnisrückmeldungen aus zentralen Vergleichsarbeiten im deutschsprachigen Raum sind eher auf die Individualebene ausgerichtet und somit der Forschung zur pädagogischen Professionalisierung zuzurechnen. Die Nutzung von evidenzbasiertem Wissen kann durch professionelle Rechenschaftslegung erklärt werden, wenn Lehrkräfte sich ihrer Profession verpflichtet fühlen. Ergebnisrückmeldungen stellen dabei eine Form der nötigen Lerngelegenheiten für die Professionalisierung von Lehrkräften dar (Darling-Hammond, 2004). Mit dem Fokus auf die Individualebene offenbart sich ein Dilemma: Einerseits scheuen einige Lehrkräfte die Beteiligung an Qualitätsentwicklungsprozessen, da sie dabei einen Teil ihrer Autonomie aufgeben müssen, andererseits scheinen gerade Kooperationen für das Gelingen dieser Prozesse notwendig (Altrichter, 2009; Altrichter & Heinrich, 2006; Opfer, Pedder & Lavicza, 2011; Peek, 2006). Derart ausgerichtete Forschung ist dadurch blind für die komplexen Gelingensprozesse und muss sich auf wahrgenommene Bedingungen beschränken, sie kann aber wesentlich genauer abbilden, woran die ideale Nutzung von evidenzbasiertem Wissen scheitert.

Grundlage der meisten aktuellen Rezeptionsstudien ist das von Helmke und Hosenfeld entwickelte Prozessmodell zur Beschreibung der pädagogischen Nutzung von Vergleichsarbeiten (Helmke, 2004; Helmke & Hosenfeld, 2005). Dieses soll hier kurz in der Version von Koch (2011) vorgestellt werden. Modifizierte Modelle findet man bei Schneewind (2007) und Groß Ophoff (2011).

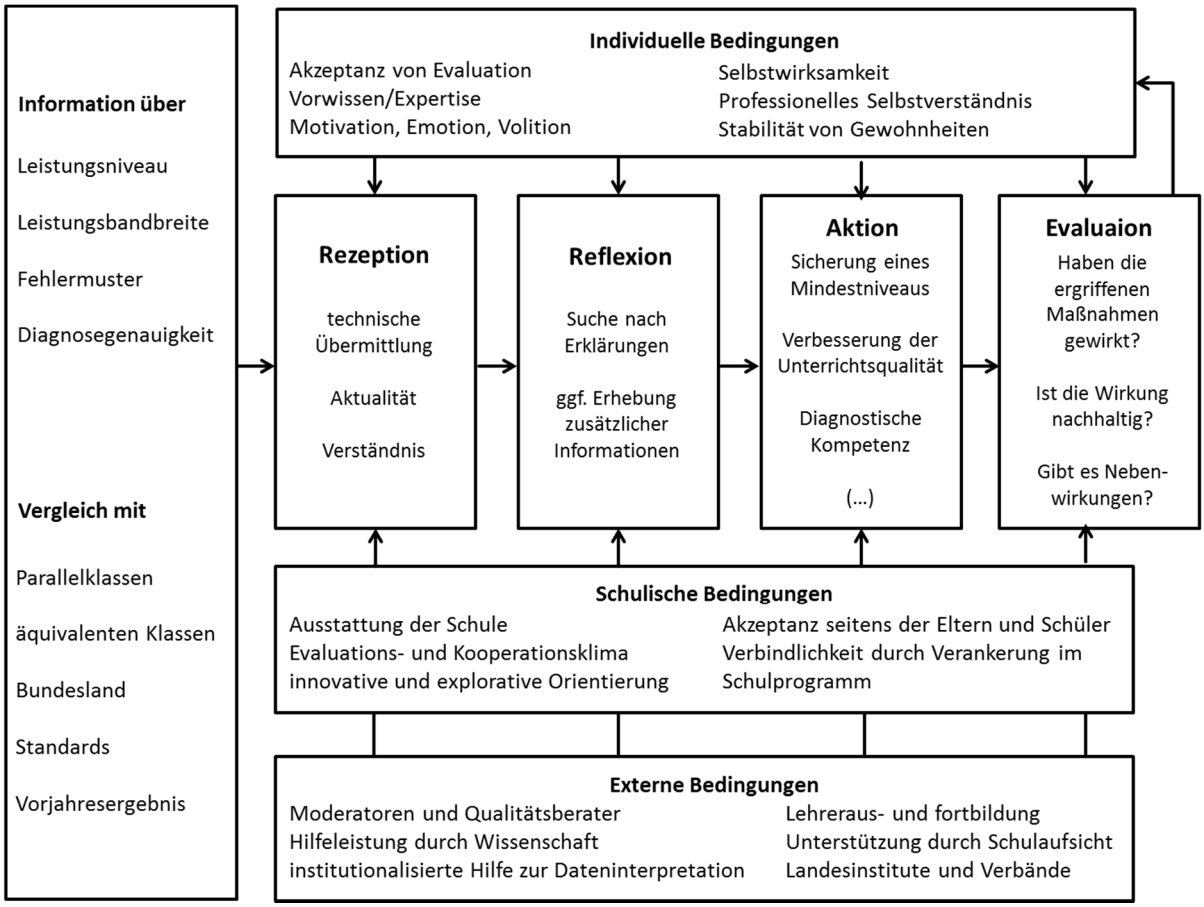


Abbildung 2.2 Rahmenmodell der pädagogischen Nutzung von Vergleichsarbeiten nach Helmke/Hosenfeld

Helmke (2004) versteht das Modell in der Tradition des Angebot-Nutzungs-Modell von Fend (2009) und sieht vier Phasen der Nutzung von Daten aus standarisierter Schulleistungsmessung vor: Es beginnt mit (1) der *Rezeption der Rückmeldungsergebnisse*. Dabei müssen die Ergebnisse bei den Lehrkräften ankommen und sie müssen möglichst aktuell sein. Unterrichtet die Lehrkraft die getesteten Schüler und Schülerinnen beispielsweise gar nicht mehr, sinkt ihre Bereitschaft, sich mit den Daten auseinanderzusetzen. Die Lehrkräfte müssen die Ergebnisse außerdem verstehen können. Es folgt (2) die *Reflexion der Ergebnisse*, wobei hauptsächlich Erklärungen für das Zustandekommen der Ergebnisse gesucht werden, aber auch zusätzliche Informationen

zugezogen werden, die die Rückmeldung nicht enthielten. Die Reflexion soll anschließend abhängig von den Ergebnissen zu einer Unterrichtsentwicklung führen, die von Helmke und Hosenfeld als (3) Aktion bezeichnet wird. Diese zielt auf die Sicherung eines Mindestniveaus und auf die Verbesserung der Unterrichtsqualität. Als Voraussetzung in dieser Phase wird eine Evaluations-, Aufgaben- und Fehlerkultur und diagnostische Kompetenz angenommen. Möglich ist hier eine Koppelung mit Projekten zur Unterrichtsqualität. Den Schluss (und theoretischen Startpunkt für den neuen Kreislauf) bildet (4) die Evaluation der Unterrichtsveränderung. In dieser Phase werden Daten gesammelt, um die gewünschte Wirksamkeit der Maßnahmen sowie Nebenwirkungen zu erfassen. Der Prozess wird in dem Modell von vier Bedingungen beeinflusst: durch die Gestaltung der angebotenen Rückmeldung, durch die schulischen, die individuellen und die externen Bedingungen. Im Sinne der Professionalisierung gibt es außerdem eine Wirkung des Prozesses auf die individuellen Bedingungen der Lehrkraft.

Umstritten ist nach empirischer Überprüfung des Modells, ob sich die vier Phasen tatsächlich als eigenständige Konstrukte identifizieren lassen. Nach Koch (2011) lassen sich die Rezeption und die Reflexion voneinander auch empirisch unterscheiden. Sie konnte zeigen, dass die Verständlichkeit eine Voraussetzung für die berichtete Intensität der Auseinandersetzung und für die den Vergleichsarbeiten zugewiesene Nützlichkeit darstellt. Auch zeigt sich eine Wirkung der Nützlichkeitseinschätzung auf die berichteten Unterrichtsmaßnahmen, die im Modell als „Aktion“ bezeichnet werden. Die Intensität der Auseinandersetzung kann diese positive Wirkung allerdings nicht aufweisen, eher führt eine besonders intensive Auseinandersetzung sogar zu weniger abgeleiteten Maßnahmen (Koch, 2011). Hosenfeld erklärt sich diesen Befund dadurch, dass bei einer intensiven Reflexion der Ergebnisse auch das Resultat stehen kann, keinen Veränderungsbedarf zu erkennen (da beispielsweise die Ergebnisse positiv sind) (Hosenfeld, 2010). Groß Ophoff (2011) empfiehlt daher den Begriff in „Prozessnutzung“ umzubenennen. Koch (2011) bilanziert, dass sich das Modell anhand der Lehrerbefragung zu VERA3 in den Jahren 2004 bis 2008 bestätigen lasse. Sie fand in einer weitergehenden Studie aber einen Zusammenhang von tatsächlichem Verständnis und tatsächlichen Auseinandersetzungsintensität (Koch, 2011). „Verständnis“ ist also anders einzuordnen als „wahrgenommene Verständlichkeit“ und ersteres von der Reflexion schwierig zu trennen. Zu dem Schluss kommt auch Groß Ophoff (2011).

Ungeachtet dessen, ob das Modell sich tatsächlich empirisch bewährt hat, zeigt es, welche komplexen Prozesse bereits auf Individualebene für eine pädagogische Nutzung vorausgesetzt werden müssen. Aus dieser Perspektive überrascht es nicht, wenn Studien (allesamt Lehrerbefragungen – Koch, Groß Ophoff, Hosenfeld & Helmke, 2006; Kühle & Peek, 2007; Maier, 2009; Tresch, 2007[46]) regelmäßig zu dem Ergebnis kommen, dass zentrale Lernstandserhebungen nur selten den prognostizierten Kreislauf angeregt haben. Trotz der bereits dargestellten Komplexität fehlt es dem Modell aber an inhaltlichen Kausalzusammenhängen (Maier, 2008a). Auch fehlt die Verbindung zu Feedbackprozessen.

[46] Ergebnisse zur Rezeptionsforschung sind in Abschnitt (4.3) dargestellt. Für eine ausführliche Übersicht siehe Koch (2011).

Feedbackprozesse (wie sie beispielsweise in der Feedback-Interventions-Theory (FIT) von Kluger und DeNisi (1996) abgebildet werden – s. (4.3)) sind zwar immer wieder Bestandteil des Theoriekonstrukts von Arbeiten zur Nutzung von Vergleichsarbeiten, werden aber darin nur lose daneben gestellt. U.a. fehlt es im Helmke-Hosenfeld-Modell an einer Zielkomponente, die für Feedbackprozesse existenziell ist (Kluger & DeNisi, 1996) (vgl. auch (4.3)). Auch Tresch (2007) hält die wahrgenommene Kongruenz von intendiertem Curriculum und Testcurriculum für eine entscheidende Voraussetzung für die datengestützte Unterrichtsentwicklung. Insgesamt stellt sich das Rahmenmodell zur pädagogischen Nutzung von Vergleichsarbeiten als ein Modell dar, welches stark auf die professionelle Eigenverantwortung der Lehrkräfte baut (auch als „Qualitäts-/Schulentwicklung durch Einsicht“ bezeichnet Böttger-Beer & Koch, 2008), während zentrale Vergleichsarbeiten trotz der Testdurchführung durch Lehrkräfte als externe Evaluation klassifiziert werden müssen (Berkemeyer & Müller, 2010). Altrichter zustimmend bleibt damit „offen, wie die individuellen Wahrnehmungs-, Handlungs- und Lernvorgänge von verschiedenen Akteuren in einem Mehrebenensystem verknüpft werden, was gerade das Grundproblem von „Systemsteuerung“ darstellt“ (Altrichter, 2010, S. 231).

2.2.5 Steuerungslogik und Wahrnehmung zentraler Vergleichsarbeiten in Deutschland

Nachdem in (2.1.) die mögliche Steuerungslogik beschrieben und dabei van Ackeren folgend drei Evaluationsansätze zu Rechenschaftslegung und Qualitätsentwicklung mit ihren Wirkungen dargestellt wurden, stellt sich an dieser Stelle die Frage, welches Evaluationsmodell der Umgang mit zentralen Vergleichsarbeiten in Deutschland folgen soll und tatsächlich folgt. Da VERA8 in der aktuellen Form erst im dritten Jahr existiert, wird dabei auch auf Erkenntnisse bzgl. vergleichbarer Studien zurückgegriffen.

Auf Seiten der schulischen Administration gibt es nach Orth (2005) klare Vorstellungen, wie mit den Ergebnissen aus zentralen Vergleichsarbeiten umgegangen werden soll. Die Ergebnisse sollen dazu dienen, an Schulen selbst etwaige Defizite zu identifizieren, und gleichzeitig sollen sich „gute Schulen“ herauskristallisieren. Gewünscht wird, dass Schulen nach dem Best-Practice-Prinzip von anderen Schulen lernen, die besonders gut abgeschnitten haben. Genauso soll dieses möglichst – und das ist wahrscheinlich noch einfacher umzusetzen – auf Klassenebene passieren. Die Analyse sollen die Schulen dabei selbst organisieren und in den Fachkonferenzen durchführen. Von einer optimalen Nutzung der zentralen Vergleichsarbeiten kann gesprochen werden, wenn die Ergebnisse in allen Fächern sachlich analysiert werden und dabei sowohl die kriteriale als auch die bezugsgruppenorientierte Perspektive als auch die Kompetenzstufenmodell berücksichtig werden und ggf. verbindliche Konsequenzen für die unterrichtliche und schulische Arbeit abgeleitet werden (Dobbelstein & Peek R., 2008). Parallel dazu muss auf administrativer Ebene nach Möglichkeiten gesucht werden, wie mit Problemschulen umgegangen werden

soll. Diese könnten in einer differenzierenden Mittelzuweisung und Fortbildungsangeboten für die Unterrichtsentwicklung bestehen (Orth, 2005). Zentrale Vergleichsarbeiten in Deutschland sind folglich erst einmal im Sinne des „Support"-Ansatzes konzipiert. Die Qualitätssicherung und -entwicklung steht im Vordergrund (Dobbelstein & Peek R., 2008). Daher soll die „Datenhoheit" bei den Einzelschulen verbleiben (Bonsen et al., 2006). Zusätzlich wird aber auch davon ausgegangen, dass Schulen vermehrt gegenüber der Öffentlichkeit Rechenschaft ablegen (Peek, 2007).

Die tatsächliche Umsetzung beschränkt die Unterstützungsleistung in den meisten Bundesländern allerdings auf die Bereitstellung der Tests und didaktischer Handreichungen sowie auf die Aufbereitung der Testergebnisse.[47] Eine Koppelung mit bestehenden fachlichen Fortbildungen existiert bisher nicht. Obwohl der Ansatz ohne starke negative Sanktionen konzipiert ist, wundert es daher nicht, dass Lehrkräfte die Vergleichsarbeiten dann wie in Baden-Württemberg als Kontrollinstrument statt als Hilfe zu professionellerem Handeln beschreiben und Ergebnisse der zentralen Vergleichsarbeiten als Legitimation für Selektionsentscheidungen nutzen (Maier, 2009). Dass es sich bei den zentralen Lernstandserhebungen um *Vergleichs*arbeiten handelt, verstärkt diesen Eindruck sicherlich.

Die Wahrnehmung zentraler Vergleichsarbeiten als Kontrollinstrument speist sich auch aus dem Umbruch, den die Neue Steuerung insgesamt für deutsche Schulen bedeutet (siehe auch van Ackeren, Heinrich & Thiel, 2013). Während zwischen schulischer Administration und Einzelschule ein Tausch von mehr Autonomie gegen mehr Rechenschaftslegung stattfindet, verlieren Lehrkräfte im Zuge dieses Umbruchs erst einmal an Privilegien. Sie werden nicht nur in ihrer Autonomie auf Unterrichtsebene beschnitten, sondern müssen auch als Einzelperson die Rechenschaft aufbringen, die die schulische Administration von der Einzelschule als Ganzes erwartet. Im Einzelnen nennen Altrichter und Heinrich: (a) Das Berufsbild der Lehrkraft wird neu definiert und verlangt neue Qualifikationen. Von Lehrkräften wird erwartet, dass sie sich innerhalb von Qualitätsmanagementsystemen einem Feedback stellen. (b) Qualitätskonzepte sehen in der Regel neue Kooperationsformen vor, und zwar sowohl während der Qualitätsentwicklung (z.B. Analyse von gewonnenen Daten in der Fachgruppe) als auch als Ergebnis der Qualitätsentwicklung (z.B. Unterrichtskooperation, Anpassung des Schulcurriculums). (c) Die Trennung von Verwaltung und Unterricht wird aufgehoben, wenn der Verwaltung für Entscheidungen Daten über den Unterricht (und seien es auch nur über Schülerleistungen vermittelte Daten) vorliegen. Schließlich droht das gewonnene Steuerungswissen auch eine Differenzierung zwischen Lehrkräften zu ermöglichen, sodass das Gleichheitsgebot ausgehebelt wird (Altrichter & Heinrich, 2006).

Im Zusammenhang mit zentralen Vergleichsarbeiten ist folglich zweierlei interessant: (1) Welche Wirkung haben zentrale Vergleichsarbeiten auf die einzelne Lehrkraft und wie geht

[47] Thüringen stellt hier eine der wenigen Ausnahmen dar und bietet seit Einführung der Kompetenztests Schulungen zur Datennutzung an Nachtigall und Jantowski (2007). Erst im Jahr 2011 ist beispielsweise damit in Nordrhein-Westfalen begonnen worden. Bereits im Jahr 2004 wurden vom damaligen Landesinstitut für Schule und Weiterbildung Informationsveranstaltungen zu Lernstand9 angeboten, diese Angebote waren aber nicht ausreichend und wurden zwischenzeitlich eingestellt.

sie mit dem Instrument um? (2) Welche Wirkung haben zentrale Vergleichsarbeiten auf die Organisation innerhalb der Einzelschule und wie wird innerhalb der Einzelschulen mit den zentralen Vergleichsarbeiten umgegangen? An dieser Stelle sollen die für Deutschland nur spärlich verfügbaren Ergebnisse zur Ebene der Einzelschulen dargestellt werden.

Die umfangreichste der Untersuchungen dazu stammt von Hartung-Beck (2009). Sie hat in einer qualitativen Studie die Organisationsstruktur zweier Gesamtschulen in Nordrhein-Westfalen bzgl. ihres Umgangs mit zentralen Vergleichsarbeiten nach Einführung von Lernstand9 untersucht. Sie ermittelte die professionellen und organisationalen Überzeugungen von neunzehn Lehrkräften und fand u.a., dass weniger die professionellen Überzeugungen denn die organisationalen darüber entscheiden, ob die Einführung von Lernstand9 begrüßt wurde oder sich Widerstand formierte. Es zeigte sich aber auch, dass Lehrkräfte nicht zwingend einen Autonomieverlust wahrnehmen (wie von Altrichter und Heinrich angenommen), sondern durch die gemeinschaftliche Aufarbeitung der Ergebnisse stattdessen einen Zugewinn an Wissen für ihre Entscheidungsgrundlage erkennen. Die Verteilung der aufgrund der Überzeugungen herausgearbeiteten Typen scheint vorbehaltlich der kleinen Stichprobe schulabhängig. Hartung-Beck identifizierte eine Kumulation von Lehrkräften mit Überzeugungen vollständig i.S. der intendierten Nutzung (z.B. rationale Erweiterung der professionellen Reflexionseben, professioneller fachlicher Diskurs, schulorganisatorische Weiterentwicklung der Zusammenarbeit) an Schule A und eine Ansammlung von Lehrkräften mit entgegengesetzten Überzeugungen (z.B. Wahrnehmung der Vergleichsarbeiten als technokratische Handlungsanweisung und fremdbestimmte Kontrolle) an Schule B. Als Wirkungsgröße stellen sich in der Studie die Beziehung der Schulleitung zu den Lehrkräften und ihr Anspruch an den Umgang mit den Vergleichsarbeiten dar.[48] Ein zentraler Führungsstil führt zu Widerstand gegenüber dem neuen Instrument, die Beziehung von Lehrkräften zur schulischen Administration (vgl. Neo-Institutionalismus) wird dabei auf die Schulleitung übertragen. Andersherum wirkt eine ausgeprägte Absprache- und Kommunikationskultur mit klar transportierten Anforderungen und Chancen für eine Affirmation gegenüber zentralen Vergleichsarbeiten (Hartung-Beck, 2009).

Auch von der Gathen (2011) fand in seiner vergleichenden Rezeptionsstudie, in der die Nutzung von Rückmeldungen aus DESI[49] und der ersten Erhebung aus Lernstand9 untersucht wurden, eine intensivere Nutzung bei denjenigen Schulen, die bereits über klare Strukturen in der Kommunikation und für die Datennutzung verfügten. Eine mögliche Erklärung für die eher geringe und oberflächliche Nutzung von durch VERA generierte Evidenzen sieht von der Gathen in der Diskrepanz zwischen wissenschaftlichem Wissen und Handlungswissen. Er vermutet eine prinzipielle Überforderung der einzelnen Lehrkraft bei der Transformation von wissenschaftlichen Evidenzen in Handlungswissen für die Praxis. Kooperationen im Sinne Professioneller Lerngemeinschaften können seiner Einschätzung nach möglicherweise Abhilfe schaffen (von der Gathen, 2011).

[48] Ähnliches berichten auch Nachtigal und Jantowski (2007).
[49] DESI steht für die Studie Deutsch-Englisch-Schülerleistungen-International.

Diemer und Kuper (2011) betrachteten durch SSL auszulösende Prozesse aus einer anderen Perspektive. Sie haben in einer qualitativen Interviewstudie an vier Schulen (ein Gymnasium und eine Gesamtschule in Berlin, ein Gymnasium und eine Realschule in Thüringen) untersucht, in welchem Umfang in den Jahren 2007 bis 2009 durch zwei Verfahren der SSL output-orientierte Steuerung initiiert wurde. Dazu beziehen sie das Begriffspaar „Konditional-/Zweckprogrammierung“ auf die Steuerungslogik der SSL im Zusammenhang mit der datengestützten Unterrichtsentwicklung. Sie sprechen von „Zweckprogrammierung“, wenn Unterrichtsentwicklung aufgrund eines Reflexionsprozesses über die Ergebnisse aus SSL resultierte, mit „Konditionalprogrammierung“ bezeichnen sie hingegen Maßnahmen der Unterrichtsentwicklung, wenn diese ohne auf Testdaten basierende Reflexionsprozesse realisiert oder angedacht wurde (Diemer & Kuper, 2011). Systematische Qualitätsentwicklung setzt Zweckprogrammierung voraus, konditionalprogrammierendes Vorgehen implizierte demgegenüber die Annahme, dass man bereits weiß, wie der angestrebte Output zu erreichen ist. Auch bei output-orientierter Steuerung kann konditionalprogrammierendes Verhalten auftreten, allerdings nur als Ausdruck der Rechenschaftslegung. Diemer und Kuper (ebenda) bilanzieren, dass Ergebnisse aus SSL eher als Prozessinformationen denn als Outputinformationen wahrgenommen werden und, wenn sie als Outputinformationenen angesehen werden, nur in wenigen Fällen Reflexionsprozesse angeregt werden. Häufiger werden die Ergebnisse lediglich als Bechmarkvergleich genutzt. Diemer und Kuper identifizieren als realisiertes Steuerungshandeln prozess-orientierte und input-orientierte Absichten: (a) von Schülerinnen und Schülern weniger gut gekonnte Inhalte zu wiederholen und stärker im Unterricht zu behandeln, (b) zu den Tests ähnliche Aufgabenformate im Unterricht zu praktizieren, (c) im Vorfeld des Tests dafür gezielt zu üben, (d) die schulinternen Bewertungsmaßstäbe anzupassen und (e) heterogene Leistungsverteilungen durch individuelle Förderung zu reduzieren auf der Prozess-Dimension sowie (f) das Schulcurriculum zu verändern, (g) didaktische Vorgaben zu vereinheitlichen und (h) Kooperations- und Kommunikationsstrukturen zu verändern. Insgesamt zeigte sich, dass prozessbezogene (und prozessnahe inhaltliche) Maßnahmen in der Regel keiner vorherigen Reflexion der Ergebnisursachen entstammen und somit keine output-orientierte Qualitätsentwicklung identifiziert werden konnte (Diemer & Kuper, 2011).

Zusammenfassend zeigt sich auf Einzelschulebene folglich ein eher unbefriedigendes Zwischenergebnis für die testbasierten Instrumente der Neuen Steuerung in Deutschland. U.a. können viele Schulleitungen ihre Schlüsselrolle für die Implementation der Instrumente einer neuen Steuerung und insbesondere einer datengestützten Schul- und Unterrichtsentwicklung nicht erfüllen (Altrichter, 2010). Dies betrifft sowohl die Organisation der Entwicklungsprozesse als auch ihre Rolle als Vermittler der Anforderungen und Möglichkeiten der Neuen Steuerung.

3 Testcoaching

Das zweite Theoriekapitel behandelt nach den Vergleichsarbeiten den anderen zentralen Begriff dieser Arbeit: das Testcoaching. Mit „Testcoaching“ kann ein Verhalten im Kontext von Prüfungen beschrieben werden, welches vermehrt durch die Begriffe „Teaching to the Test“ oder „Teaching for the Test“ charakterisiert wird. Diese in den scheinbaren Synonymen ausgedrückte negative Konnotation wird dem Komplex Testcoaching nicht gerecht und verkürzt den Begriffsinhalt. Indem die drei Begriffe gleichgesetzt werden, scheinen eine generelle Vorbereitung auf Test bzw. Prüfungen, ein durch einen Test initiiertes Lernen und die Auswahl von Lern- und Vorbereitungsinhalten nicht von einer Stoffverknappung unterscheidbar, die durch eingeführte (zentrale) Tests erzwungen wurden. Gleichsam werden positive Effekte einer Testvorbereitung wie das Vertrautmachen mit Testsituationen als unerwünschte Nebeneffekte deklariert.[50] Dem stehen insbesondere die Implementation- und die Innovationsfunktionen von Vergleichsarbeiten gegenüber, aber auch Maßnahmen für eine höhere Testvalidität werden dadurch unterdrückt.

Das folgende Kapitel steht somit unter der Zielsetzung, den Begriff “Testcoaching“ differenzierter zu erfassen. Dazu wird zuerst eine testtheoretische Betrachtung für die Item-Response-Theorie und die allgemeine Messfehlertheorie vorgenommen. Dabei ist der Begriff “Beta Abilities“ als eine Testkompetenz (im Sinne von Fertigkeit und Fähigkeit) zentral. Ohne die Existenz einer Testkompetenz wäre ein entscheidender Teil von Testcoaching nicht möglich. Anschließend erst kann der Versuch einer Definition unternommen werden. Qualitätsbetrachtungen (i. S. wertneutraler Klassifikationen) zeigen Parallelen und Abgrenzungen zu Nachhilfe und gewöhnlichen Lern-Übungszeit. Testcoaching wird hier als eine bestimmte Form der Unterrichtsqualität dargestellt, bei der die (Lern-)Intention klassifizierend wirkt, nicht der Lernerfolg. Anschließend werden hingegen Effekte von Testcoaching auf die Leistungsmessung bei zentralen Tests in den U.S.A und bei PISA 2003 dargestellt. In welcher Weise verschiedene Variationen von Testcoaching effektive Messwertsteigerungen erreichen können, wird ihm Rahmen dieser Arbeit allerdings nicht untersucht. Der Blick über den Tellerrand dient lediglich, um eine der wesentlichen Wirkungen von Testcoaching, die Veränderung des Messwertes bei Tests, zu veranschaulichen. Den Schluss dieses Kapitels bilden daher die Möglichkeiten zu Testcoaching bei VERA 8 in Nordrhein-Westfalen sowie der diese Arbeit motivierende Befund aus der Befragung von 18 Lehrkräften zu ihrem Vorbereitungsverhalten aus dem Jahr 2008.

[50] Entsprechende Gleichsetzungen finden sich etwa in deutschen Artikeln von Bonsen & von der Gathen (2004) oder Lind (2009), in englischen Artikeln beispielsweise bei Au (2007) oder Amrein & Berliner (2002).

3.1 Eine Skizze der testtheoretischen Grundlagen

Testcoaching betrifft folglich alle drei Hauptgütekriterien (Objektivität, Reliabilität und Validität). Um zu verstehen, worin die Gefahren (oder auch Möglichkeiten) von Testcoaching liegen können, ist ein gewisses Grundverständnis über die testtheoretischen Grundlagen der verwendeten Tests nötig. Im folgenden Abschnitt wird Testcoaching unter der Item-Response-Theorie (IRT) und der allgemeinen Messfehlertheorie (aMT) (auch Klassische Testtheorie genannt) betrachtet. Es wird sich zeigen, dass Testcoaching aus dem Blickwinkel der IRT ein Validitätsproblem, aus Sicht der aMT ein Reliabilitätsproblem darstellt.

Die Objektivitätsgefährdung ist als Durchführungsobjektivität vorwiegend durch die Möglichkeit von eindeutig unerwünschten Verhalten der Testleitung gefährdet. Dies geschieht beispielsweise, indem einigen Probanden (Schülerinnen und Schülern) zusätzliche Informationen gegeben werden, die das Testergebnis beeinflussen. Der einfachste Fall ist das Helfen bei der Bearbeitung von Aufgaben durch den Testleiter (die Lehrkraft). Andersherum stellt ein zu geringer Informationsgrad der Testleitung (der durchführenden Lehrkraft) als Konsequenz eigener mangelhafter Testvorbereitung die Durchführungsobjektivität in Frage. Hier kann es beispielsweise aufgrund nicht vollständig gelesener Durchführungsbestimmungen zur falschen Anweisung der Probanden (Schülerinnen und Schüler) kommen, bestimmte Hilfsmittel nicht zu verwenden. Selbstverständlich sind auch Defizite im Sinne der Auswertungs- und Interpretationsobjektivität aufgrund nicht ausreichender Vorbereitung der Testleitung trotz eindeutiger Auswertungsanleitungen denkbar. Diese Probleme sind unabhängig von einer Testtheorie offensichtlich, betreffen aber nicht die Kernidee von Testcoaching.

Traditionell kann man unter einem (psychologischen) Test ein psychologisches Experiment verstehen, welches mit dem Ziel durchgeführt wird, vergleichende Aussagen über Personen zu erhalten (Rost, 2004). Auch zentrale Vergleichsarbeiten sind als psychologisches Experiment aufzufassen, bei der die zu messende Variable nicht identisch mit den erhobenen Antwortmustern ist. Benötigt wird folglich eine Testtheorie, die ein beobachtetes Antwortverhalten (manifeste Variable) zu einer zu erfassenden Kompetenz (als latente Variable operationalisiert) in Beziehung setzen kann. Im Rahmen von „VERA 8“ bedient man sich dazu ebenfalls der IRT[51]. Innerhalb der IRT gibt es verschiedene statistische Modelle, die zur Testkonstruktion herangezogen werden können. Die Aufgaben im Projekt „VERA 8“ sind für ein dichotomes Rasch-Modell, bei dem die Antworten nur den Kategorien „gelöst“ oder „nicht gelöst“ zugeordnet werden können, kalibriert und skaliert (Greve, 2011). Auf

[51] Korrekterweise müsste bei VERA3 und VERA8 von einem zweistufigen Modell gesprochen werden, denn nach dem Schluss von den Antwortmustern auf die jeweilige Kompetenzausprägung der Schüler wird anschließend im Idealfall von der Kompetenzausprägung auf die Qualität des Unterrichts geschlossen.

Länderebene kann aber auch ein „Partial-Credit-Modell“[52] genutzt werden, wenn zusätzlich Teillösungen abgebildet werden sollen.

Das (genauer: dichotome) Rasch-Modell (Einparameter-Logistisches Modell) zeichnet sich durch eine monoton steigende, punktsymmetrische Itemfunktion aus, die an ihrem Wendepunkt quasi linear verläuft und sich für die Lösungswahrscheinlichkeit null und eins parallel zur X-Achse anschmiegt. Das wichtigste Charakteristikum ist der *eineindeutige* Itemparameter (i.S.v. eindeutig bis auf Translationen in Richtung der Argument-Achse) für alle Items, sodass die Itemfunktionen des Tests alle parallel auf der X-Achse verschoben sind (Rost, 2004). Sowohl der Personenparameter als auch der Itemparameter werden auf derselben Skala abgetragen (Moosbrugger, 2008a). Diese Skala ordnet die Personenfähigkeiten und Itemschwierigkeiten bezüglich einer normierten Sozialnorm, welche auf Grundlage der Anzahl der von einer Person gelösten Items und der Anzahl der ein Item lösenden Personen bestimmt wird (Moosbrugger, 2008a; Rost, 2004). Die Modellgleichung des Raschmodells gibt nach Interpretation des Autors die Wahrscheinlichkeit dafür an, dass ein Item der Schwierigkeit X von einer Person mit der Personenfähigkeit Y gelöst wird.[53] Das Argument der Funktion besteht aus der Differenz der Personenfähigkeit und der Itemschwierigkeit. Die Itemschwierigkeit ist im Rasch-Modell für jede Person aus der Stichprobe identisch (Steyer & Eid, 2001).[54]

Die IRT wird im deutschen Sprachraum auch als probabilistische Testtheorie bezeichnet, weil häufig auf Modelle zurückgegriffen wird, die über eine Wahrscheinlichkeitskomponente verfügen (Bühner, 2010).[55] Bei der Zuordnung der manifesten Variablen zu latenten Variablen wird angenommen, dass die direkte Zuordnung (‚Proband besitzt Kompetenz‘ drückt sich in ‚Lösung des Items‘ aus) nur mit einer gewissen Wahrscheinlichkeit zutrifft. Diese Wahrscheinlichkeitsmodellierung umfasst bereits zufällige Messfehler, sofern sie aus dem Schluss von den Lösungsmustern auf das Ausmaß der latente Variable resultieren (Carstensen, Knoll, Rost & Prenzel, 2004). Zu diesen zufälligen Messfehlern können durch einen Rateparameter auch systematische, der Aufgabenanlage entspringende Messfehler berücksichtigt werden.[56]

Gleichzeitig wird über die probabilistische Testtheorie eine starke Restriktion über den Ursprung des gezeigten Antwortverhaltens vorgenommen. Das Raschmodell und verwandte

[52] Das Partial-Credit-Modell beruht auf dem Rasch-Modell und addiert die Einträge des mehrdimensionalen Personenvariablen-Vektors zu einem eindimensionalen Personenvariablen-Vektor wie im Rasch-Modell benötigt (Rost, 2004).

[53] Nach einer alternativen Interpretation gibt der Wahrscheinlichkeitswert an, wie viele Testpersonen das Item lösen (Wainer, 2010).

[54] >>Hier müssen noch Formeln ergänzt werden.<<

[55] Gemeint ist die Manifestierung der Kompetenz im Antwortverhalten. In der aMT wird diese stets unterstellt. Gleichwohl besitzt besitzen auch Modelle der aMT zumindest eine Wahrscheinlichkeitskomponente, denn Messfehler werden als zufällig angesehen.

[56] Handelt es sich bei der Aufgabe beispielsweise um eine Aufgabe, bei der die Lösung aus vier vorgegebenen Antwortmöglichkeiten ausgewählt werden muss, wird bei rein zufälliger Auswahl durchschnittlich jedes vierte Mal die richtige Antwort ausgewählt. Wird dies bei der Schätzung der Itemschwierigkeit nicht berücksichtigt, wird diese systematisch unterschätzt.

Modelle setzen das latente Konstrukt, welches der manifesten Variablen (dem Antwortverhalten) zugrunde liegt, als ein eindimensionales (aufgefasst als Itemhomogenität) (Bühner, 2010). Dieses Verständnis wird für die Leistungsmessung in Schulleistungsstudien in zwei Richtungen kritisiert: Einerseits müssen zur Beantwortung der Testitems häufig neben der intendierten Kompetenz weitere Fähigkeiten herangezogen werden, sodass das zu messende Konstrukt mehrdimensional scheint, andererseits wird die Theorie vertreten, die in den (internationalen) Schulleistungsstudien unterschiedenen Kompetenzen ließen sich auf einen g-Faktor zurückführen (Rindermann, 2006).

Auch die Idee des Testcoachings beruht auf der Vorstellung, dass die latenten Konstrukte mehrdimensional sind und durch einen gemeinsamen Faktor beeinflusst werden. Anders als bei beiden Richtungen der Kritik von Rindermann und Kollegen an (internationalen) Schulleistungsstudien wird das latente Konstrukt allerdings weder in verwandte Konstrukte zerlegt, die ebenfalls unabhängig von Testsituationen eine Existenzberechtigung an sich reklamieren können, noch wird ein der zu messenden Kompetenz übergeordneter g-Faktor unterstellt. Die hier postulierte Testkompetenz (als Testfertigkeit oder auch Testfähigkeit) wird gerade nur durch Testsituationen relevant.

Eine zentrale Voraussetzung der Itemhomogenität in der IRT ist die lokale stochastische Unabhängigkeit. Der Begriff setzt sich aus der stochastischen Unabhängigkeit und der als lokal bezeichneten Fixierung des zu messenden Personenparameters zusammen. Inhaltlich sind mit lokaler stochastischer Unabhängigkeit zwei Vorstellungen verbunden, erstens besteht zwischen den Items keine logische Abhängigkeit (wie es beispielsweise bei Filterfragen der Fall ist) und zweitens existiert keine Abhängigkeit durch Reihungs- und Positionseffekte im Sinne reaktionskontingenten Veränderungen (beispielsweise Lernen durch Einsicht oder Verstärkungslernen) (Rost, 2004). Lokale stochastische Unabhängigkeit ist die Voraussetzung für Itemhomogenität und Eindimensionalität des zu messenden Konstrukts (Bühner, 2010). Lokale stochastische Unabhängigkeit impliziert die Vorstellung von einfacher lokaler Unabhängigkeit, die in der aMT vorausgesetzt wird. Lokale Unabhängigkeit meint lediglich die Nicht-Korrelation von Items eines Tests, die alle denselben Personenparameter messen, wenn dieser fixiert wird. Bestehen bei konstanter Ausprägung des zu messenden Personenparameters weiterhin (Residual-)Korrelationen zwischen einzelnen oder allen Items, treten systematische Messfehler auf (Bühner, 2010). Mathematisch bezeichnet die einfache lokale Unabhängigkeit die *paarweise* stochastische Unabhängigkeit. Diese lässt sich wie folgt definieren:

Eine Familie $\{A_i , i \in I\}$ von Ereignissen heißt *paarweise [stochastisch] unabhängig*, wenn für alle $i \neq j$ die Ereignisse A_i , A_j unabhängig sind, also $P(A_i \cap A_j) = P(Ai) \cdot P(Aj)$ (Krengel, 2005. S.26).

Die lokale stochastische Unabhängigkeit kann, muss aber nicht durch Testcoaching verletzt werden. Die Eindimensionalität des zu messenden Konstrukts ist nicht gefährdet, wenn mehrere psychologische Prozesse beteiligt sind, dabei aber stets gleichzeitig auftreten (Bühner, 2010). Stetiges Auftreten drückt sich in der lokalen stochastischen Unabhängigkeit

dadurch aus, dass alle beteiligten Prozesse gleichzeitig fixiert werden können, wenn die Personenfähigkeit konstant gehalten wird. Dies kann quasi auch für die von Bond als *Beta Abilities* (s.u. Bond, 1993) bezeichnete Testfertigkeit gelten. Dies setzt allerdings ein dichotomes Auftreten voraus und teilt die Probanden in zwei nicht miteinander vergleichbare Gruppen (Probanden, die die nötige Testfertigkeit besitzen, und Probanden, die die nötige Testfertigkeit nicht besitzen).[57] Andersherum ist die lokale stochastische Unabhängigkeit sofort verletzt, wenn die Beta Abilities statt einer Testfertigkeit eine Testfähigkeit darstellen und unabhängig von der zu messenden Kompetenz einen Einfluss auf den wahren Testwert besitzen.[58] Es geht dabei sowohl um die Möglichkeit, dass eine Testkompetenz die eigentlich zu messende Kompetenz ersetzt und dadurch bessere Ergebnisse in den Lösungsmustern abgebildet werden als die eigentlich zu messende Kompetenz erwarten lassen sollte, als auch um die Möglichkeit, dass ein gewisser Grad an Testkompetenz Voraussetzung ist, um den Grad der eigentlich zu messenden Kompetenz adäquat in den Lösungsmustern abbilden zu können. Die Testkompetenz ist dann keine einseitig störende Größe, deren Vorkommen bei den einzelnen Probanden so gering wie möglich sein sollte. Viel mehr wird ein gewisser Grad an Testkompetenz für eine ungestörte Abbildung der zu messenden Kompetenz in der Testsituation unabdingbar.

Von der aMT hingegen werden zufällige und systematische Messfehler behandelt, die der Messsituation bzw. dem Messvorgang zuzurechnen sind. Während traditionell in der Messfehlertheorie aber angenommen wird, der beobachtete Wert einer Messung setze sich nur aus wahrem Wert und Messfehler zusammen (Rost, 2004), erweitert Bond das Modell für Messungen von Messgegenständen, die durch Lernen beeinflussbar sind, noch um die *Beta Abilities* (Bond, 1993). Auch Steyer und Eid unterscheiden hierbei zwischen unsystematischen äußeren oder inneren Einflüssen und Übungs- und Transfereffekten (Steyer et al., 2001).

Die zentralen Annahmen der aMT sind das Existenzaxiom (mit dem Erwartungswert einer Zufallsvariable existiert ein wahrer Testwert), das Verknüpfungsaxoim (der Messwert setzt sich aus dem wahren Testwert und dem Fehlerwert zusammen, wobei der Erwartungswert des Fehler null beträgt) und das Unabhängigkeitsaxiom (wahrer Testwert und Fehler sind unkorreliert) (Moosbrugger, 2008b). Zur Bestimmung des Fehlerintervalls, mit dem von Messwert auf den wahren Wert geschlossen wird, über die Reliabilität (definiert als der Quotient aus Varianz des wahren Testwerts und Varianz des Messwerts) bzw. die Reliabilität allgemein wird außerdem vorausgesetzt, dass die Messfehler zwischen verschiedenen Items unkorreliert sind (Steyer et al., 2001).

[57] Ein Vertrautsein gleichermaßen aller Testprobanden ist folglich in jedem Fall die Voraussetzung für die Nutzung von IRT-Modellen.

[58] Die in Abs, (3.4) dargestellten Ergebnisse zu Testcoaching im amerikanischen und israelischen Raum lassen vermuten, dass es sich bei den Beta Abilities wahrscheinlich um eine Testfähigkeit handelt. Zur Klärung bedarf es allerdings spezifischere Untersuchungen.

In dem Fall gilt

$$\begin{aligned} Var(X) &= Var(T+E) \\ &= Var(T) + Var(E) + 2 \cdot Cov(T,E) \\ &= Var(T) + Var(E), \end{aligned}$$

mit Var(X): Varianz von X, Cov(X,Y): Kovarianz von X und Y, X: Messwert.

Aus Sicht der aMT können die Beta Abilities als zweiter Bestandteil des wahren Testwerts oder als systematischer – aber nicht für alle Probanden gleicher - Messfehler aufgefasst. Werden die Beta Abilities als zweiter Bestandteil des wahren Testwerts betrachtet, folgt

mit T=A+B, A: Kompetenz, B: Beta Abilities

$$\begin{aligned} Var(X) &= Var(A+B+E) \\ &= Var(A+B) + 2Cov(A+B, E) + Var(E) \\ &= Var(A) + Var(B) + 2 \cdot Cov(A,B) + Var(E) + 2 \cdot [Cov(A,E) + Cov(B,E)] \end{aligned}$$

und mit den Axiomen auch für die Beta Abilities

$$Var(X) = Var(A) + Var(B) + 2 \cdot Cov(A,B) + Var(E).$$

Unter der Annahme der Unkorreliertheit von wahrem Testwert der Kompetenz und Beta Abilities bleibt somit die Varianz der Beta Abilities als zusätzliche, nicht eindeutig bestimmbare Größe. Der wahre Testwert ist unter dieser Betrachtung nur noch in Kombination mit anderen wahren Testwerten interpretierbar.

Im zweiten Fall ist der Messfehler zweier Items resultierend nicht mehr als unkorreliert annehmbar und die Gleichung

$$\begin{aligned} Var(X) &= Var(T+E) \\ &= Var(T) + Var(E) + 2 \cdot Cov(T,E) \end{aligned}$$

ist nicht auf die Gleichung

$$Var(X) = Var(T) + Var(E)$$

zurückführbar. Daraus resultierend kann die Varianzzerlegung nicht genutzt werden, um vom Messwert auf den wahren Wert zu schließen.

Beide formale Interpretationen von Beta Abilities innerhalb der aMT führen somit zu Schwierigkeiten in der Interpretation der Messwerte. Inhaltlich kann Testcoaching in einer ersten Annäherung als Pre-Test im Sinne einer Test-Retest-Situation aufgefasst werden. Mit dem Test soll der Erfolg eines Lernprozesses gemessen werden, dessen Abschlussdatum nicht mit dem Testdatum zusammenfällt, sondern vor den Beginn des Testcoachings anzusiedeln ist. Systematische, aber ungleichmäßige Messfehler im Zusammenhang mit der Test-Retest-Reliabilität werden immer wieder diskutiert. Lienert und Raatz weisen auf Effekte hin, die durch Wiedererkennen von Aufgaben entstehen und zu konsistenten

Antworten führen könnten. Sie bezeichnen die durch Memorierung von Aufgabeninhalten bei Nutzung der Test-Retest-Methode zu erwartende höhere Reliabilität als Scheinreliabilität, die u.a. durch zu einfachere Aufgaben und kurze Zeitabstände zwischen dem Test und der Testwiederholung befördert werden (Lienert & Raatz, 1998). Auch Schermelleh-Engel und Werner sehen die Gefahr von Übungseffekte und Wissenszuwachs in Leistungstest. Diese sind entsprechend nicht als systematisch gleichmäßig anzunehmen (Schermelleh-Engel & Werner, 2008).

Zusammenfassend kann die Gefährdung der Objektivität, der Reliabilität und – auch für die aMT zumindest resultierend – der Validität durch Maßnahmen, die Lehrkräfte und Schüler und Schülerinnen unter dem Titel der Testvorbereitung durchführen, unter dem Blickwinkel beider Testtheorien festgehalten werden. Im folgenden Abschnitt werden daran anschließend die verschiedenen Qualitätsausprägungen von Testcoaching analysiert und eine Definition gegeben.

3.2 Testcoaching: Definition, Test Wiseness und Herangehensweisen[59]

3.2.1 Eine Definition für Testcoaching

Von Testcoaching als vollständiger Begriff wird selten gesprochen. Gebräuchlicher sind die scheinbar synonymen Begriffe “Teaching to the Test“ oder “Teaching for the Test“. Ersteres betont eine Vorbereitung, die durch die Ankündigung eines Tests initiiert wird (Sjuts, 2007). Im Mittelpunkt steht formal die Steuerungswirkung von zentralen Tests, die aber häufig negativ gesehen und mit der Dopingpraxis des Sports gleichgesetzt wird (Popham, 2001). Letzteres meint die Vermittlung von Lerninhalten aus der Intention heraus, gute Messergebnisse zu erzielen. Nach von der Gathen führt dies unter Umständen zur ausschließlichen Vermittlung von vorher bekannten Testinhalten und begünstigt somit eine Curriculumsreduktion. Lehrkräfte reduzierten die Unterrichtsinhalte auf diejenigen Inhalte, von denen sie annähmen, dass diese Testgegenstand seien (von der Gathen, 2011). Wenn weiter unten die Inhalte von Testcoaching erläutert werden, wird deutlich werden, dass diese beiden Begriffe nur einen Teilaspekt von Testcoaching umfassen. Nichtsdestoweniger muss eine Definition für Testcoaching folglich zwei Dingen Rechnung tragen: Von welcher Art ist die Vorbereitung? Welches Ziel wird mit der Vorbereitung angestrebt?

Mit “Testcoaching“ könnte in einer ersten Näherung ein Vorbereitungsverhalten im Kontext von Prüfungen beschrieben werden. Diese simple Definition entspricht weitestgehend der

[59] Die Abschnitte 3.2 bis 3.4.1 sind fast vollständig aus Hahn (2008) übernommen.

Fragestellung der Testcoaching-Forschung zu standarisierten Tests im anglo-amerikanischen Bereich (z.B. Allalouf & Ben-Shakhar, 1998; Messick, 1981; Powers, 1985). Laut Allalouf und Ben-Shakhar sucht die Testcoaching-Forschung nach der Antwort, ob es möglich ist, in Intelligenztests und schulischen Fähigkeitstests durch Interventionen höhere Messwerte zu erzielen. Dahinter steht die Frage nach der prognostischen Validität von Tests. Wenn Allalouf und Ben-Shakhar unter Testcoaching alles Verhalten verstehen, dass der Vorbereitung auf Tests dient (Allalouf et al., 1998), wird zwar der Steuerungsaspekt von Testcoaching angesprochen, eine Abgrenzung von normalen Lern- und Übungsphasen unterbleibt allerdings.

Die im vorherigen Abschnitt auf Bond (1993) zurückgehende mit Beta Abilities bezeichnete Testkompetenz trägt dieser Abgrenzung besser Rechnung. Testcoaching ist demnach als Vorbereitungsverhalten definiert, welches die als Beta Abilities bezeichnete Kompetenz erhöhen soll. Mit einer derartigen Definition beraubt man sich allerdings der Möglichkeit, zwischen bewusst vorgenommenen Handlungen und Handlungen, die zufällig und unbeabsichtigt geschehen, zu unterscheiden. Zwischen diesen Handlungen zu trennen scheint aber in so weit sinnvoll, als dass mit Testcoaching auch immer die Frage nach Kontrolle von Faktoren verbunden ist und zufällige und unbeabsichtigte Handlungen anders zu handhaben sind.

Testcoaching sei daher zunächst jede bewusst vorgenommene Tätigkeit, die Beta Abilities zu erhöhen.

Neben der Differenzierung von Beta Abilities (s.u.) liefert auch die Klassifikation des Vorbereitungsverhaltens durch Allalouf und Ben-Shakhar (1998) in drei Herangehensweisen (Familiarity Approach, Content Approach und Test Wiseness Approach) bei Testcoaching einen wichtigen Hinweis: Während die Aktivitäten zum Familiarity Approach zum und Test Wiseness Approach auf die Beta Abilities abzielen, fallen Aktivitäten im Sinne eines Content Approach nicht unter diese Definition. Üblicherweise nehmen (zusätzliche) Übungsphasen vor summativ verwendeten Klassenarbeite, Klausuren und Tests aber einen wesentlichen Platz ein. Es scheint folglich sinnvoll, einen Kompetenzzuwachs, der allein oder vorwiegend für einen höheren Testscore angestrebt wurde und der sich auf für den Test relevante Teilkompetenzen beschränkt, von einem allgemeinen Kompetenzzuwachs zu unterscheiden, der allgemein auf Lern- und Übungseffekte zurückzuführen ist. Unsere Arbeitsdefinition des Testcoachings soll deswegen wie folgt lauten:

Testcoaching ist jede bewusst vorgenommene Handlung, um die Beta Abilities zu erhöhen oder einen Kompetenzzuwachs zu erreichen, der allein oder vorwiegend für einen höheren Testscore angestrebt wurde und der sich auf für den Test relevante Teilkompetenzen beschränkt.

Obwohl eine Testvorbereitung anschaulich in der Zeit direkt vor dem Test und auch in einem relativ engen Rahmen stattfindet, enthält die Definition bewusst weder eine Bedingung für den Zeitpunkt noch für den Zeitraum der Vorbereitung. Dieser Umstand ist der besonderen

Stellung des Steuerungsaspekts von angekündigten Tests geschuldet. Eine monatelange Vorbereitung ist bei weit vorher angekündigten Tests gleichsam denkbar wie eine regelmäßige, aber nicht ununterbrochene Vorbereitung. Die Definition wird stringenter, wenn sie stattdessen nur den Inhalt der Vorbereitung abdeckt und diesen mit der Intention der Vorbereitung verknüpft.

3.2.2 Test Wiseness

Test Wiseness wird von Millman, Bishop und Ebel als „a subjects's capacity to utilize the characteristics and formats of the test and test-taking situation to receive a high score" definiert (Millman, Bishop, Ebel, 1965, S.707 nach Bond, 1993, S.431). Test Wiseness ist also eine Kompetenz, aber auch Emotionen wie Testangst, die hemmende Wirkung auf das Testergebnis haben können, und das Selbstkonzepts gehören zur erweiterten Test Wiseness und auch die generelle Intelligenz der Probanden spielt wie auch bei der eigentlich zu messenden Kompetenz hinein. Wichtig zu bemerken ist die Ausweitung des Konzepts von Test Wiseness auf Fähigkeits- sowie auf Persönlichkeitstests: Für Persönlichkeitstests geht es nicht um „High Scores", sondern um sozial erwünschte Antworten.

Nach Pike (1978) und Millman, Bishop und Ebel (1965) lässt sich Test Wiseness in *General Test Wiseness* und *Test-specific Test Wiseness* einteilen (Pike 1978; Millman, et al., 1965, beide nach Bond, 1993). General Test Wiseness umfasst Vertrautheit mit der Bearbeitung von Tests, eine geschickte Zeiteinteilung, Rate-Strategien usw. Test-specific Test Wiseness meint ähnliche Elemente, welche schon unter die Kategorie General Test Wiseness eingeordnet wurden. Hier bezieht sich die Test Wiseness aber auf einen bestimmten Test (Pike, 1978, nach Bond, 1993). Test-specific Test Wiseness setzt Wissen über den konkreten Test voraus. Dies kann beispielsweise durch Beispielaufgaben oder Aufgaben als alten Testungen bereitgestellt sein.

Betrachtet man die Fülle der an verschiedener Stelle gegebenen Test-Wiseness-Strategien, scheint es verständlich, wenn Flippo, Becker und Wark behaupten, dass es durchaus auch als kognitive Leistung angesehen werden muss, diese Test-Wiseness-Strategien anzuwenden, und noch mehr, diese zu entwickeln. Der fehlende Zusammenhang von Anwenden von Test-Wiseness-Strategien und tatsächlich gebrauchter Test-Bearbeitungszeit unterstreicht diese Vermutung noch einmal und betont auch die Bedeutung allgemeiner Intelligenz bei der Anwendung solcher Strategien (Brunner et al., 2007; Flippo et al., 2000). Eine systematische Untersuchung von Test Wiseness steht aber noch aus (Bond & Harman, 1995).

Einige Beispiele seien hier gegeben (Bond, 1993; Flippo et al., 2000):

Halte dich nicht zu lange an einer Frage auf!

Mache dich mit dem Antwortformat vertraut!

Ziehe alle Antworten in Betracht, bevor du dich für eine entscheidest!

Lies die Instruktionen und Fragen genau!

Rate bei Multiple-Choice-Fragen, wenn es keine Minuspunkte für falsche Antworten gibt oder du ausreichend viele Antworten ausschließen kannst!

Bearbeite zuerst die Fragen, bei denen du dir sicher bist!

Notiere dir spontane Einfälle!

Achte auf die grammatikalische Einschränkung der möglichen Antworten!

Die erste Idee ist meist die beste!

3.2.3 Herangehensweisen

Nach Allalouf und Ben-Shakhar lassen sich drei Herangehensweisen bei Testcoaching unterscheiden (Allalouf et al., 1998; auch Baydar, 1990):

(1) Familiarity Approach: Die Teilnehmer machen sich mit den Elementen des Tests vertraut.[60] Dazu werden nach Möglichkeit frühere Testversionen analysiert und man setzt sich mit den Testinstruktionen, der Testzeit und dem Frage-Antwort-Format auseinander. Auch wird mit den Materialen die Testsituation unter authentischen Bedingungen simuliert. Dieses Verfahren dient auch dazu, um hemmende Emotionen wie Testangst zu reduzieren oder das Selbstkonzept zu stärken, und lässt Probanden die Testzeit effektiver nutzen. Beta Abilities, die in dieser Form erweitert werden, haben einen positiven Effekt auf die Validität der vorbereiteten Tests. Probanden werden nicht durch Unkenntnis des Testformats oder durch die besondere psychische Situation davon abgehalten, den Test unter ihren Möglichkeiten zu bearbeiten (Baydar, 1990).

(2) Content Approach: Die Teilnehmer bereiten sich intensiv auf die Inhalte vor. Beispielsweise werden für einen Test zum Problemlösen vorwiegend Problemlöse-Aufgaben bearbeitet. Diese Art der Vorbereitung beeinflusst die Validität des Tests nicht zwingend, da Elemente eingeübt werden, die vollständig Teil der zu messenden Kompetenz sind. Problematisch wird es allerdings, wenn der Test Auskunft über den Unterrichtserfolg geben

[60] Anders als Brunner und Mitarbeiter (2007) dies sehen, zählt ein Pretest als Kontroll-Bedingung in einem Experiment nicht unter Testcoaching oder Testvorbereitung dieser Art, weil jener Pretest nur ungewollte Effekte hat und damit nicht unserer Definition entspricht.

soll und nicht das ganze intendierte Curriculum erfasst wird. Hier besteht die Gefahr, geringerer Validität und überschätzter Testergebnisse (von der Gathen, 2011).[61]

(3) Test Wiseness Approach[62]: Die Teilnehmer beschäftigen sich mit Test-Wiseness-Strategien. Dabei sind sowohl General Test Wiseness als auch Test-specific Test Wiseness angesprochen. Test Wiseness-Strategien werden häufig als Tricks angesehen (Bond, 1993) und besitzen damit eine negative Konnotation, da die Vermittlung von Test Wiseness-Strategien dazu dient, einen höheren Messwert im Test zu erlangen als dies die eigentlich zu messende Kompetenz zulässt. Abgesehen von den hier beispielhaft genannten Strategien (v), (viii) und (ix) sind die Strategien aber häufig aus dem Bereich des Problemlösens entliehen und werden – wie bei den Lernstandserhebungen im Projekt „Lernstand 8“ – teilweise auch in den Testheftinstruktionen explizit genannt. Abhängig von der konkreten Test Wiseness-Strategie kann dadurch die Validität des Tests gefährdet, aber auch erhöht werden.

3.2.4 Testcoaching als Teil der Unterrichtsqualität

Die „Qualität“ einer Sache kann zweierlei ausdrücken: erstens die wertneutrale Beschaffenheit dieser Sache und zweitens die Ausprägung jener bereits vorausgesetzten Beschaffenheit. Entsprechend kann die Frage nach Unterrichtsqualität auf die Art und Weise zielen, inwieweit diese normativen Vorstellungen genügen, und sie kann auf die Effizienz der vorgefundenen Art und Weise gerichtet sein. Aus der zweiten Möglichkeit zu fragen ergibt sich die Frage nach gutem Unterricht als Unterricht, der die besten Leistungen ermöglicht, die richtigen Einstellungen, Handlungen und Werte vermittelt (Ditton, 2009). Es ist Helmke Recht zu geben, wenn er Unterrichtsqualität primär auf das fachliche Lernen ausgerichtet sieht (Helmke, 2009). Seine Forderung, die Unterrichtsqualität müsse sich an empirisch messbaren Erträgen manifestieren, bedingt aber gerade auch die oben beschriebenen Voraussetzungen (u.a. Vertrautheit mit den Testformaten, keine Testangst, ein elaboriertes Selbstkonzept, Problemlösestrategien) für den Umgang mit Tests. Anders als Helmke dies darstellt, sind Schlüsselkompetenzen und affektive, emotionale und motivationale Orientierungen vom Aufbau intelligenten Wissens nicht zu trennen, da Lernen nicht von Überprüfen zu trennen ist. Dies lässt Testcoaching als Art und Weise der Vorbereitung auf Tests zu einer Frage der Unterrichtsqualität werden.

[61] Während dies früher grundsätzlich ein Problem von Lernstand 8 war, da aus testökonomischen Gründen nur zyklisch Teilkompetenzen der Fächer Deutsch, Englisch und Mathematik getestet wurden, bilden die durch das IQB für VERA 8 entwickelten Mathematik-Aufgaben alle Leitideen ab. Eine auf bestimmte Teilkompetenzen zielende Vorbereitung ist zwar weiterhin möglich, kann aber nicht als auf die Steigerung des Testscores zielend angesehen werden. Für Deutsch und die erste Fremdsprache bei VERA 8 und für Deutsch und Mathematik bei VERA 3 hingegen ist es weiterhin üblich, nur Teil-Prozesskompetenzen zu testen. Diese werden auf der Internetseite des IBQ im Vorfeld bekannt gegeben.

[62] Brunner und Mitarbeiter benutzen in Anlehnung an Allalouf und Ben-Shakhar Test Wiseness in einer anderen Art als wir dies nach Bond machen.

Testcoaching-Programme in den U.S.A. und vergleichbare Angebote in anderen Staaten finden meistens außerunterrichtlich, oft sogar eben außerschulisch statt. Sie sind vergleichbar mit den Angeboten des Nachhilfesektors in Deutschland. Beide Angebotsformen zielen vollständig oder vorwiegend auf Leistungssteigerungen in schulischen Prüfungssituationen ab und sind abhängig vom Professionalisierungsrad der Anbieter stark selektiv (Baydar, 1990; Flippo et al., 2000; Rudolph, 2009). Über die tatsächliche Gestaltung der Testcoachingstunden existiert wenig wissenschaftlich-systematisches Wissen. Wie Baydar bemerkt, steht bisher nur die Frage im Mittelpunkt der Forschung, in welchem Umfang Effekte auf den Testscore von außerunterrichtlichen (und oft auch außerschulischen) Testcoaching-Programmen zu beobachten sind (Baydar, 1990). Die spezielle Qualität des Vorbereitungsverhaltens bleibt dabei außen vor.

In der vorgeschalteten qualitativen Erhebung (s. 3.5) hat sich die Vorbereitung auf Lernstand 8 als Teil des regulären Unterrichts gezeigt. Mit Blick auf die auch formative Funktion der zentralen Vergleichsarbeiten (diagnostische Funktion), aber auch auf die prognostische Funktion bei VERA 3, die höchste Anforderungen an die Validität der Tests stellen, und mit Blick auf die positiven Effekte von Familiarity Approach und bestimmten Elementen des Test Wiseness Approach auf das Selbstkonzept und andere affektive, emotionale und motivationale Orientierungen muss Testcoaching auch als Teil der Unterrichtsqualität verstanden werden. Die in Vorbereitung auf die zentralen Vergleichsarbeiten thematisierten Elemente können als Test Wiseness auch die Leistungsmessung an anderer Stelle valider und effektiver machen[63]. Die Einordnung als Unterrichtsqualität gilt ebenso für die als Content Approach zu klassifizierenden Übungsphasen aufgrund der implementierenden und innovierenden Funktion. Gerade auch hier können situationsübergreifende Fachinhalte gelernt werden.

Testcoaching kann aber – wie in den vorherigen Abschnitten erläutert – auch negative Effekte haben, die Validität der Instrumente gefährden und wertvolle Unterrichtszeit verschwenden. Unter anderem gilt es das richtige Maß zwischen zu viel und zu wenig Vorbereitung zu finden (Powers, 1998). Dadurch ist die Vorbereitung nicht bloß als ein allgemeiner Aspekt eine Frage der Unterrichtsinhalte, sondern es ist eine genaue Betrachtung von Art und Umfang der Vorbereitung entscheidend. Ein möglicher Grund für Messwertverzerrungen durch Testcoaching sind (zu) komplexe Test-Wiseness-Strategien. Es ist nicht klar, dass alle Schüler diese Test Wiseness-Strategien in gleicher Weise erlernen können. Viel mehr deuten Studien darauf hin, dass kognitives Potenzial sehr großen Einfluss darauf hat, ob dieser Teil der Test-Kompetenz erlernt werden kann (Brunner et al., 2007)

[63] Zu diesem Schluss kommen auch Allalouf und Ben-Shakhar (1998) sowie Powers (1985) in zwei Untersuchungen.

3.3 Forschungsstand zu Testcoaching bei PISA, SAT und GRE

Wenn Becker bemerkt, die Forschung zu Testcoaching habe eine lange, zum jetzigen Zeitpunkt über fünfzigjährige Tradition (Becker, 1990), so gilt dies vorwiegend für Testcoaching als Vorbereitung zu High-Stake-Tests, speziell im Zusammenhang mit dem *SAT* und *Graduate Record Examination (GRE)*. Weder Internetsuchmaschinen noch einschlägige Datenbanken für wissenschaftliche Artikel liefern nennenswerte Einträge zu Low-Stake-Test-Studien. Einzig eine Studie zur Effektivität von Testcoaching zu PISA 2003 von Brunner und Mitarbeitern (Brunner et al., 2007) ist zu finden. Diese Forschungsergebnisse sind nur schwierig auf die Vergleichsarbeiten zu übertragen. Dies liegt einmal daran, dass wir den Begriff des Teststellenwerts differenzierter in zwei Dimensionen betrachten wollen (vgl. (2.1.2)) und Vergleichsarbeiten nicht unbedingt als High-Low-Stake oder gar High-High-Stake-Tests gelten können. Forschungsergebnisse von High-Stake-Test-Studien sind wahrscheinlich grundsätzlich nicht auf Low-Stake-Tests verallgemeinerbar (Brunner et al., 2007). Zweitens fehlt es in Deutschland an einer Testkultur, wie man sie beispielsweise in den U.S.A vorfindet (Bonsen & von der Gathen, 2004). Dadurch fehlt es möglicherweise auch gleichsam bei deutschen Lehrkräften an Wissen zu Test Wiseness und Erfahrung mit Testcoaching.

Trotzdem sollen auch die Befunde zum Testcoaching und zur Vorbereitung auf High-Stake-Tests hier dargestellt werden, weil sie zumindest einen gewissen Rahmen skizzieren können, in dem sich auch Ergebnisse zu Low-Stake-Test-Studien bewegen sollten. Dabei sind drei Fragenkomplexe denkbar: (1) Welche Effekte hat Testcoaching auf die Messwerte des Test und welche Maßnahmen bringen eine Zunahme des Messwerts? (2) In welchem Rahmen werden Testcoaching-Programme genutzt und aus welchen Beweggründen geschieht dies? (3) Welchen Einfluss haben die Effekte von Testcoaching auf den Aussagewert von Tests und wie sollte damit umgegangen werden? – Da bei Studien zur Gefährdung der Testvalidität durch Testcoaching keine ausreichend Differenzierung nach verschiedenen Herangehensweisen vorgenommen wurde, werden im Folgenden nur die Ergebnisse zu den ersten beiden Fragen berichtet.

3.3.1 Allgemeine Effekte von Testcoaching

Studien zum Testcoaching für High-Stake-Tests gibt es beinahe unzählige. Allerdings leiden diese häufig an sehr kleinen Stichproben und nicht wenige weisen wissenschaftliche Mängel auf. Daher wird vermehrt auf Meta-Analysen zurückgegriffen, um die allgemeinen Effekte von Testcoaching auf Messwerte zu beschreiben (Becker, 1990), (Bond, 1993; DerSimonian & Laird, 1983; Flippo et al., 2000; Kulik et al., 1984; Messick, 1981).

Kulik und Mitarbeiter resümieren, dass in 25 von 38 der von ihnen analysierten Studien Testcoaching einen signifikant positiven Effekt aufwies und 13-mal der Effekt positiv, aber nicht signifikant war. Sie berichten von einem Effekt in der Größenordnung von über vierzig Prozent, durchschnittlich von bis zu einem Drittel einer Standardabweichung. Allerdings gehe ein großer Anteil daran auf Pretest-Effekte (und andere unbewusste, auch den Kontrollgruppen zukommende Effekte) zurück (Kulik et al., 1984).[64] Andere Meta-Analysen kommen zu geringeren Effektgrößen von vierundzwanzig Prozent oder gar nur neunzehn (SAT-M) bzw. neun Prozent (SAT-V), wenn sich die Analyse auf Studien mit Kontrollgruppen-Design beschränkt (Becker, 1990; Bond, 1993).

Brunner und Mitarbeiter berichten in ihrer Studie zur Effektivität von Testcoaching bei PISA 2003 von keinem Effekt bei der Lesekompetenz, für die Mathematik-Kompetenz stellten sie aber immerhin Effekte durch Testcoaching und Pretest von vier bis vierundzwanzig Punkten fest (Brunner et al., 2007).

Testcoaching scheint grundsätzlich effektiver für Mathematik-Tests als für Tests aus dem Bereich der Sprache (Lesekompetenz und bei PISA 2003 und SAT-verbal). Hier ist besonders auffällig, dass Testcoaching für den Bereich der Sprache teilweise sogar negative Effekte zu haben scheint, welches für Mathematik-Tests nicht zu beobachten ist (Becker, 1990), (Brunner et al., 2007; Flippo et al., 2000; Kulik et al., 1984).

Durch die Meta-Analysen geht aber jegliche Information über die Qualität von Testvorbereitung und Testcoaching verloren, sodass die Wirkung von Testcoaching erst einmal nur wenig differenziert betrachtet werden kann. Zumindest in einer Studie zum Testcoaching für den SAT, die mit über viertausend Testpersonen eine ausreichend große Stichprobe vorweisen kann, zeigte sich deutlich, wie unterschiedlich Testcoaching wirkt. Die Teilnehmer von Testcoaching-Programmen erreichten im Posttest einen Messwert, der von null bis zweihundert Prozent der Standardabweichung betrug[65] (Powers, 1999). Dies kann einmal auf eine geringe Reliabilität des verwendeten SATs deuten, es kann aber genauso als Indiz dafür verstanden werden, welche unterschiedliche Qualität die Varianten der Testvorbereitung aufweisen.

Bunting und Mooney und auch Messick und Jungeblut schließen durch Meta-Analysen, dass ein Testcoaching-Programm mindestens drei Stunden dauern muss, um überhaupt positive Effekte zu erzielen (Bunting & Mooney, 2001; Messick, 1981). Flippo, Becker und Wark geben für die effektivsten Testcoaching-Programme[66] eine Zeit zwischen sechs und neun

[64] Kulik und Mitarbeiter (1984) halten die Größenordnung, die sie in ihrer Studie entdeckten, für SAT-Scores vernachlässigbar, was bei einer Verteilung der Scores zwischen 200 und 800 Punkten durchaus zutreffend ist. Für Studien wie PISA 2003 wäre ein Testcoaching-Effekt in dieser Größenordnung allerdings verheerend. Hätte Deutschland dank Testcoaching bei PISA 2003 in der Mathematik-Kompetenz eine Drittelstandardabweichung mehr Testpunkte erzielt, gehörte es in Mathematik zu der Spitzengruppe.

[65] Durchschnittlich betrugen die Testcoaching-Zuwächse 21% der Standardabweichung für den SAT-V und 34% der Standardabweichung für den SAT-M.

[66] Untersucht wurden nur kurzzeitig durchgeführte Programme. Programme, die über mehrere Monate laufen, wurden ausgeklammert.

Stunden an und Kulik und Mitarbeiter konnten bei einige Stunden länger dauernden Programmen keine größere Effektivität mehr erkennen (Flippo et al., 2000; Kulik et al., 1984). In diesem Zusammenhang ist auch die Folgerung von Messick und Jungeblut zu verstehen, dass motivierte Teilnehmer größeren Nutzen aus Testcoaching-Programmen ziehen als unmotivierte (Messick, 1981). Bedenkt man den geringen zeitlichen Umfang, den man für ein Vertrautmachen mit dem Multiple-Choice-Format voraussichtlich benötigt, scheinen die Effekte eher auf eine tatsächliche Auseinandersetzung mit den Testinhalten zurückführbar.

Bezüglich der Frage, ob Test Wiseness-Strategien Erfolg versprechend sind, kommen manche Studien zu keiner positiven Antwort, andere hingegen deuten auf positive Effekte. Tatsächlich systematisch untersucht hat dies Baydar. In ihrer Simulationsstudie wurde Testcoaching nur in Form von Familiarity Approach und als Test Wiseness Approach berücksichtigt. Der hier errechnete Effekt von Testcoaching-Programmen ist nach Baydar eher gering und im Vergleich zu den Selektionseffekten für die Teilnahme an diesen Programmen durch den sozioökonomischen Status vernachlässigbar (Baydar, 1990). Flippo hält es trotzdem beziehungsweise gerade deswegen für eine pädagogische Pflicht der Schulen, diese Strategien zu thematisieren, um niemanden zu benachteiligen (Flippo et al., 2000).

3.3.2 Motivation zu Testcoaching und seine Verbreitung

Schon im Abschnitt zur High-Stake- und Low-Stake-Tests ist deutlich geworden, welcher Stellenwert der Motivation bei Tests zukommt. Die Motivation spielt aber nicht nur eine Rolle, während der Test bearbeitet wird. Powers fand heraus, dass – wie man vielleicht auch erwartet hat – Teilnehmer an Testcoaching-Programmen für den *Graduate Record Examination (GRE)* ein größeres Interesse an einem hohen Score besitzen. Aber das Interesse beschränkte sich nicht nur auf den Messwert des Tests. Sie waren auch mehr an einem höheren Bildungsabschluss interessiert, beschäftigten sich häufiger mit Aktivitäten über den Studienverlaufsplan hinaus und strebten häufiger die Aufnahme in eine Graduate School an (Powers, 1985). Baydar entdeckte unter SAT-Testpersonen, dass Teilnehmer von Testcoaching-Programmen häufiger mit ihrem Abschneiden beim ersten SAT-Versuch (berechtigt) unzufrieden waren und durchschnittlich geringere Scores erreichten als ungecoachte Testpersonen. Gleichzeitig stammten die gecoachten Testpersonen überproportional aus Familien mit überdurchschnittlichem Einkommen und Bildungsstand (Baydar, 1990). Powers und Rock wiesen nach, dass die Teilnehmer an Testcoaching-Programmen häufiger beim ersten Versuch im SAT an Nervosität litten als Testteilnehmer, die auf Coaching verzichteten (Powers, 1999). Teilnehmer an Testcoaching-Programmen sind also einerseits an höheren Bildungsabschlüssen interessiert und andererseits befürchten sie mehr, diese Bildungsziele nicht zu erreichen, wenn sie sich nicht besonders vorbereiten.

Wie aber sieht diese Vorbereitung in den meisten Fällen aus? Auf die Frage nach den verwendeten Mitteln, um sich auf den SAT vorzubereiten, gaben diejenigen Befragten, die auch angaben, gecoacht worden zu sein, in einer Studie aus den Jahren 1995/96 am häufigsten an, ein Buch zur allgemeinen Testvorbereitung und ein Buch, welches sich speziell mit dem SAT beschäftigt, gelesen zu haben. Gut die Hälfte hatte Beispielaufgaben bearbeitet und nur ein Drittel nutzte die Unterrichtsmaterialien der vergangenen Jahre. Systematisches Testcoaching erlebten 39% in der Klasse, 19% in Zusatzkursen der Schule und 15% außerhalb auf privater Basis. Damit unterschieden sich die Befragten, die angaben, nicht gecoacht worden zu sein, nur unwesentlich. Nur ein Buch zur allgemeinen Testvorbereitung lasen bei den ungecoachten Befragten nicht einmal halb so viele (Powers, 1999). Möglicherweise stimmen die Zahlen für außerschulische Testcoaching-Programme heute allerdings nicht mehr. Parallel zur Zunahme des Anteils der Schülerinnen und Schüler, die in Deutschland Nachhilfe bekommen (Haag, 2006), ist auch in diesem Bereich eine Zunahme denkbar. In einem gewissen Umfang konnte dies schon im Verlaufe der Achtzigerjahre beobachtet werden (Baydar, 1990). Powers hingegen konnte die von ihm erhobenen Anteile zwischen 1985/86 und 1995/96 jedoch bestätigen (Powers, 1988, 1998).

In jedem Fall zeigen die Befunde, dass die Schulen beim Testcoaching eine führende Rolle spielen und private Angebote (zu diesem Zeitpunkt) nur von wenigen genutzt werden (vielleicht genutzt werden können). Bei einer Befragung unter 508 weiterführenden Schulen in den U.S.A. (Rücklaufquote 68%) in den Jahren 1986/87 und erneut 1995/96 unter 576 (Rücklaufquote 60%) ergab sich, dass ungefähr die Hälfte der Schulen Testcoaching-Programme angeboten hat, die teilweise an den normalen Unterricht angeschlossen waren und teilweise unabhängig von ihm stattfanden (Powers, 1988, 1998). Dies deckt sich also zumindest mit der in den Siebzigerjahren durchgeführten Schülerbefragung. Bei fast allen Schulen, die entsprechende Programme anboten, geschah dies mit dem Ziel, die Schülerinnen und Schüler mit dem SAT vertrauter zu machen und die Scores im SAT-M und SAT-V zu verbessern. Auch relativ häufig sollte das Selbstkonzept aufgebaut und Prüfungsangst reduziert (letzteres aber mit abnehmender Tendenz) werden. Die Hälfte dieser Schulen wollte darüber hinaus Test-Wiseness-Strategien vermitteln. Verbale oder mathematische Kompetenz zu erweitern, war bei der Erhebung 1985/86 nur im Fokus von 36% bzw. gar nur 28% der Programme, zehn Jahre später gaben dies mit aber schon 52% und 43% deutlich mehr Schulen als Ziel an. Umgesetzt werden sollten diese Ziele vor allem durch Bearbeiten eines auf den SAT vorbereitendes Buchs, Testsimulationen[67] und von den Schulen selbst entwickeltes Material (Powers, 1988, 1998).

[67] Buch und Testbögen wurden vom College Board, eine amerikanische gemeinnützige Organisation als Zusammenschluss der u.s.-amerikanischen Hochschulen und mit der Durchführung des SAT beauftragt, herausgegeben.

3.4 Testcoaching bei VERA8

3.4.1 Möglichkeiten und Instrumente zum Testcoaching bei VERA8

Grundsätzlich können für VERA 8 (und auch VERA 3) wie auch bei U.S.-amerikanischen standarisierten High-Stake-Tests zwei Quellen für Testinformationen und Vorbereitungsmaterialien herangezogen werden: die für die Konstruktion, formale Durchführung und Auswertung verantwortlichen staatlichen Institutionen und im Bereich des Schulbildungswesens privatwirtschaftlich agierende Verlage.

Informationen und Materialien der staatlichen Testverantwortlichen

Ähnlich dem College Board beziehungsweise dem Educational Testing Service (ETS) bieten das IQB und das MSW Informationen über die Ziele, den Ablauf, die Testinhalte und –konstruktion. Diese werden allerdings nicht in einem speziellen Vorbereitungsmanual angeboten, sondern sind als Hinweise sowie durch Beispielaufgaben und Aufgaben aus früheren Testungen verfügbar. Explizit wird auf der Internetseite des MSW auf eine mögliche Vorbereitung im Vorfeld der VERA 8-Erhebung eingegangen. Dort heißt es in den Hinweisen für Lehrkräfte zur Vorbereitung von Schülerinnen und Schülern:

„Lernstandserhebungen beziehen sich im Unterschied zu Klassenarbeiten nicht auf den unmittelbar vorher im Unterricht erarbeiteten Stoff. Ein kurzfristiges Üben von Aufgaben, ist deshalb nicht erforderlich, weil mit der Lernstandserhebung zurückgemeldet werden soll, welche Kompetenzen die Klasse bzw. der Kurs längerfristig erworben hat.

Das Ziel ist es, mithilfe dieser Ergebnisse Hinweise für die Unterrichtsentwicklung abzuleiten. Gezieltes Üben von Aufgaben kann diesem Ziel nicht gerecht werden.

Die Aufgabenformate der Lernstandserhebungen können sich von denen der Klassenarbeiten unterscheiden. Die Schülerinnen und Schüler können aber über den Ablauf sowie die Anforderungen der Lernstandserhebungen informiert und mit den Aufgabenformaten vertraut gemacht werden (Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, 2010a).“

Ähnlich steht es in den Hinweisen für Eltern und Schüler:

„Der Aufbau allgemeiner Kompetenzen ist ein langfristiger Prozess, der über Jahre hinweg in guten Lernarrangements erfolgt, die auch Alltagsbezüge, das Vernetzen von Inhalten, regelmäßiges Wiederholen von Grundlagen, gegenseitiges Erklären, Zusammenstellen und Umarbeiten von Übersichten, eigene Schreibprodukte usw. beinhalten.

Ein gezieltes Trainieren von Testaufgaben kann diesen Kompetenzaufbau nicht ersetzen. Es verfälscht aber das Ergebnis und den Nutzen der Lernstandserhebung, weil hierbei vielleicht kurzfristig Wissen oder Fertigkeiten eingebracht werden, die als dauerhafte Kompetenz noch gar nicht verfügbar sind.

Es ist dagegen sinnvoll, die Schülerinnen und Schüler auf die ungewohnten Aufgabenformate vorzubereiten (Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, 2010b).“

Während das IQB eine kurzfristige Vorbereitung als lediglich nicht sinnvoll bezeichnet(Institut zur Qualitätsentwicklung im Bildungswesen), warnen beide Hinweise auf der Internetseite des MSW davor, die Testvalidität durch Üben/Trainieren von Aufgaben zu gefährden und sich über den Kompetenzaufbau durch den Erwerb von kurzfristigem Wissen und Fertigkeiten zu täuschen.[68] Die bereitgestellten Beispielaufgaben und alten Testinhalte sollen ausschließlich Familiarity Approach-Maßnahmen (und der für staatliche Organisationen in Demokratien üblichen Rechenschaftslegung) dienen. Eine Vorbereitung im Sinne eines Test-Wiseness Approach und eines Content Approach sind folglich nicht erwünscht. Dies steht im Widerspruch zu der implementierenden Funktion, nach der durch die eingesetzten Aufgaben die Inhalte von Bildungsstandards und Kernlehrplänen illustriert werden sollen. Auch der innovierenden Funktion ist der Einsatz von alten Testaufgaben und Beispielaufgaben wahrscheinlich zuträglich: Jakobs sieht im Lernen mit entsprechenden Testaufgaben eine sinnvolle Möglichkeit, die angestrebten Kompetenzen zu erwerben (Jakobs, 2008). Nach der Theorie der Tacit knowing View kommt dem Lernen von implizitem Wissen an Aufgaben eine entscheidende Rolle zu (Neuweg, 2008). Der Umfang der veröffentlichten Aufgaben bietet aber sehr wohl die Möglichkeit einer mehrstündigen Übungsphase. Für die Fächern werden zwischen 79 (Mathematik – VERA 8) und sechs Aufgaben (Deutsch - VERA 3) als Beispielaufgaben bereitgestellt (Stand: 24.07.2011). Durch die Schwerpunktsetzung in den Fächern Deutsch und Englisch und ab 2012 auch in Französisch ist sogar eine auf Teilkompetenzen beschränkte Vorbereitung möglich.

Privatwirtschaftlich agierende Verlage im Schulbildungswesen:

Ausgangspunkt für die Forschungsbefunde des nachfolgenden Abschnitts sind Internetrecherchen mit der Suchmaschine *Google.de* unter den Begriffen „Lernstandserhebung“, „VERA 8“, „Vergleichsarbeiten + Verlag“ und „Vorbereitung auf VERA 8“ sowie darauf aufbauende Recherchen innerhalb der Internetpräsenzen der Verlage *Cornelsen Verlag, Duden-Schulbuchverlag, Ernst Klett Verlag, Schulbuchverlag Schroedel, Stark Verlagsgesellschaft* und *School-Scout-Verlagsgesellschaft*[69] im Juni 2008 und im Februar 2010.

Seit 2007 werden von den untersuchten Verlagen[70] spezielle Hefte für die Vorbereitung auf Lernstand 8 und seit dem Jahr 2008 länderübergreifende Hefte für die Vorbereitung auf Lernstandserhebungen/VERA 8 herausgebracht. Diese Vorbereitungshefte existieren für die Fächer Deutsch, Englisch (jeweils teilweise mit Audio-CD) und Mathematik und werden von

[68] Bemerkenswert an den Formulierungen ist allerdings, dass das Üben mit Aufgaben in den Hinweisen für Lehrkräfte als „nicht erforderlich“ bezeichnet wird, das Äquivalent in den Hinweisen für Erziehungsberechtigte und Schülerinnen und Schüler hingegen fehlt. Hier wird lediglich auf mögliche negative Konsequenzen durch das Trainieren von Aufgaben hingewiesen.

[69] Die School-Scout-Verlagsgesellschaft ist ein vom Verband deutscher Schulbuchverlage seit 2002 anerkanntes Mitglied. Der Verlag bietet eigene Dokumente und Werke anderer Verlage zum kostenpflichtigen Download an.

[70] Der Duden-Schulbuchverlag (Teil der Cornelsen Verlagsgruppe seit 2009) und der Schulbuchverlag Schroedel haben 2008 Vorbereitungshefte angeboten, 2010 hingen nicht mehr.

einigen Verlagen für die Schwierigkeitsdifferenzierungen der VERA 8-Testhefte „Grundanforderungen (A)“, „Mittlere Anforderungen (B)“ und „Erweiterte Anforderungen (C)“ angeboten. Sie beinhalten in der Regel den VERA 8-Test nachempfundene Testaufgaben und teilweise Erläuterungen zu den einzelnen Kompetenzbereichen. Stellenweise finden sich Test-Wiseness-Strategien in den Heften. Die Ansprache in den Heften lässt zwei verschiedene Intentionen erkennen: Einige Hefte richten sich an Schülerinnen und Schüler als Lernende und thematisieren dabei die diagnostische und Kompetenzentwicklung unterstützende Funktion von Vergleichsarbeiten, andere Hefte sprechen Schüler als Leistende an und stellen den Einfluss der Ergebnisse aus den VERA 8-Tests auf die Zeugnisnote in den Vordergrund.

Möglichkeiten zum Testcoaching

Durch das Angebot der Verlage, vor allem aber durch die auf der Internetseite bereitgestellten Informationen, ist es möglich, alle drei Arten von Testcoaching (vgl. (3.2.3)) auch für VERA 8 umzusetzen:

(1) Familiarity Approach: Es ist möglich einen Pretest schreiben zu lassen oder auch nur einzelne Aufgaben zu lösen, um sich mit der Testkonstruktion, der Testzeit und dem Frage-Antwort-Formaten von VERA 8 vertraut zu machen und die Testsituation unter authentischen Bedingungen zu simulieren.

(2) Content Approach: Die innovative und die implementierende Funktion drücken aus, dass grundsätzlich eine intensive Vorbereitung auf die Inhalte gewünscht ist. Es ist aber auch eine kurzfristige konzentrierte Vorbereitung auf die Inhalte möglich, ohne dass diese Inhalte in der Form in den Unterricht des ganzen Schuljahres oder gar der ganzen Schulzeit der Klassen 5 bis 8 wie gewünscht eingeflossen sind. Besonders durch die Schwerpunktesetzung in der Erhebung in Deutsch und den Fremdsprachen ist diese auf eben jene Schwerpunkte konzentrierte und beschränkte Vorbereitung möglich. Dies steht natürlich im Widerspruch zur durch die zentralen Vergleichsarbeiten angestrebten Unterrichtsentwicklung.

(3) Test Wiseness Approach: Test Wiseness-Strategien sind in den Hinweisen auf den Internetseiten zu VERA 8 und in den Vorbereitungsheften zu finden. Dabei handelt es sich auch in den Vorbereitungsheften um Problemlöse- und Bearbeitungsstrategien. „Tricks“, die eine Lösung der Aufgaben ohne den Besitz der entsprechenden Kompetenz ermöglichen, lassen sich nicht finden.

3.4.2 Ergebnisse einer qualitativen Studie zum Testcoaching bei Lernstand8

Das Untersuchungsdesign

Zentraler Untersuchungsgegenstand der qualitativen Interviewstudie (Hahn, 2008) war die von Lehrkräften durchgeführte Vorbereitung auf die zentralen Lernstandserhebungen des Jahres 2008 in Nordrhein-Westfalen. In so genannten halbstrukturierten, offenen Interviews wurden das Vorbereitungsverhalten und Gründe dafür sowie Einstellungen zu Lernstand 8 erhoben. Die Fragen deckten alle drei Formen des Testcoachings ab und berücksichtigten alle theoretisch aus der vorherigen Dokumentenanalyse von Internetseiten und gedruckten Vorbereitungsmaterialien. Neben Fragen nach der Vorbereitung im Unterricht wurden die Lehrkräfte auch nach dem Umfang der Vorbereitung, Veränderungen der Vorbereitung im Vergleich zu früheren Jahren und nach einer außerschulischen, individuellen Vorbereitung der Schülerinnen und Schüler befragt. Außerdem wurde gefragt, ob die Lehrerinnen und Lehrer ihrer Meinung nach übereinstimmend mit den Kernlehrplänen unterrichteten, welchen Stellenwert die Lernstandserhebungen für die beteiligten Schüler und Schülerinnen, Lehrkräfte und Erziehungsberechtigten haben und welche Funktionen sie den Lernstandserhebungen zuschreiben. Die Vorbereitung im Unterricht und das individuelle außerschulische Vorbereitungsverhalten der Schüler wurden vergleichend auch in sieben ähnlich angelegten Schülerinterviews erhoben. Die Beschreibungen der Vorbereitungsphasen im Unterricht entsprachen dabei den Angaben der Lehrkräfte.

Bei den befragten Lehrkräften handelt es sich um achtzehn Lehrerinnen und Lehrer von zwei Dortmunder und einem Wuppertaler Gymnasium, einer Dortmunder Realschule und einer Bochumer Gesamtschule. Zwei der befragten Schülerinnen und Schüler besuchen eines der Dortmunder Gymnasium, fünf Schülerinnen und Schüler das Wuppertaler Gymnasium.

Die aufgezeichneten Interviews wurden transkribiert und die Interviewprotokolle wurden anschließend einer qualitativen Inhaltsanalyse unterzogen. Dazu wurden elf Kategorien gebildet wie „Content Approach“ oder „Stellenwert für Lehrerinnen und Lehrer“ und mit passenden Kodiermöglichkeiten charakterisiert (vgl. auch *Ergebnisse*). Die elf Kategorien und ihre sie charakterisierenden Kodiermöglichkeiten wurden in einem Zusammenspiel aus theoretischen Vorüberlegungen und Probekodierungen einiger Transkriptionsprotokolle entwickelt. Der Wert für die Interraterreliabilität lag knapp über .70 und war damit noch im zufrieden stellenden Bereich (Bos, 1989).

Ausgewählte Ergebnisse

In sechzehn der achtzehn Fachlehrerinterview-Protokolle konnten Testcoaching-Maßnahmen identifiziert werden. In jedem dieser Protokolle wurde mindestens eine Maßnahme kodiert, die unter „Familiarity Approach“ eingeordnet wurde. Die am häufigsten gefundenen Maßnahmen waren mit elf bzw. zehn Kodierungen „Einsatz von

Vorbereitungsheften“ und „Training mit alten Testaufgaben“ gefolgt von „Trainieren der Aufgabenstellung“ und „Trainieren der Antwortformate“ (jeweils acht Kodierungen).

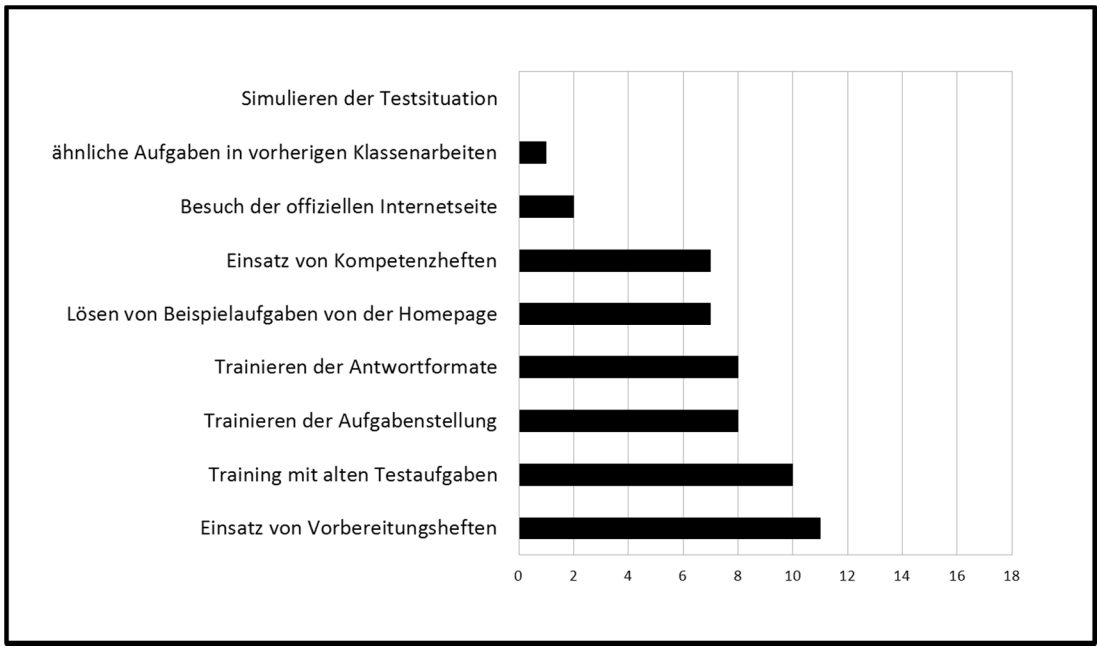


Abbildung 3.1: Kodierhäufigkeit in der Kategorie „Familiarity Approach“

Die meisten Lehrkräfte, die überhaupt Testcoaching durchführten, nutzten drei oder vier der möglichen Maßnahmen (sechsmal bzw. fünfmal). Von den sechzehn coachenden Lehrern nutzte nur ein Lehrer weder Vorbereitungshefte noch Kompetenzhefte.

Fünfzehnmal wurden auch Kodierungen zum Bereich „Content Approach“ vorgenommen. Dabei zeigte sich eine besondere Berücksichtigung der jeweiligen Schwerpunkte der zentralen Lernstandserhebungen im Jahr 2008 für alle drei Fächer (vgl. Abb. 3.2).[71] Allerdings wurde in Deutsch und in Englisch auch von einigen Lehrkräften die Lesekompetenz im Rahmen der kurzfristigen Vorbereitung gezielt gefördert.

Test-Wiseness-Strategien wurden immerhin noch in neun Protokollen kodiert. Nur dreimal wurde allerdings eine Strategie vermittelt, die eine richtige Lösung ermöglicht ohne die eigentlich zu messende Kompetenz zu vermitteln.

[71] Im Jahr 2008 wurden für die zentralen Lernstandserhebungen in Mathematik Aufgaben mit der Prozesskompetenz „Werkzeuge“ als Schwerpunkt eingesetzt. In Deutsch und Englisch lag der Schwerpunkt auf dem Verfassen kleiner Texte.

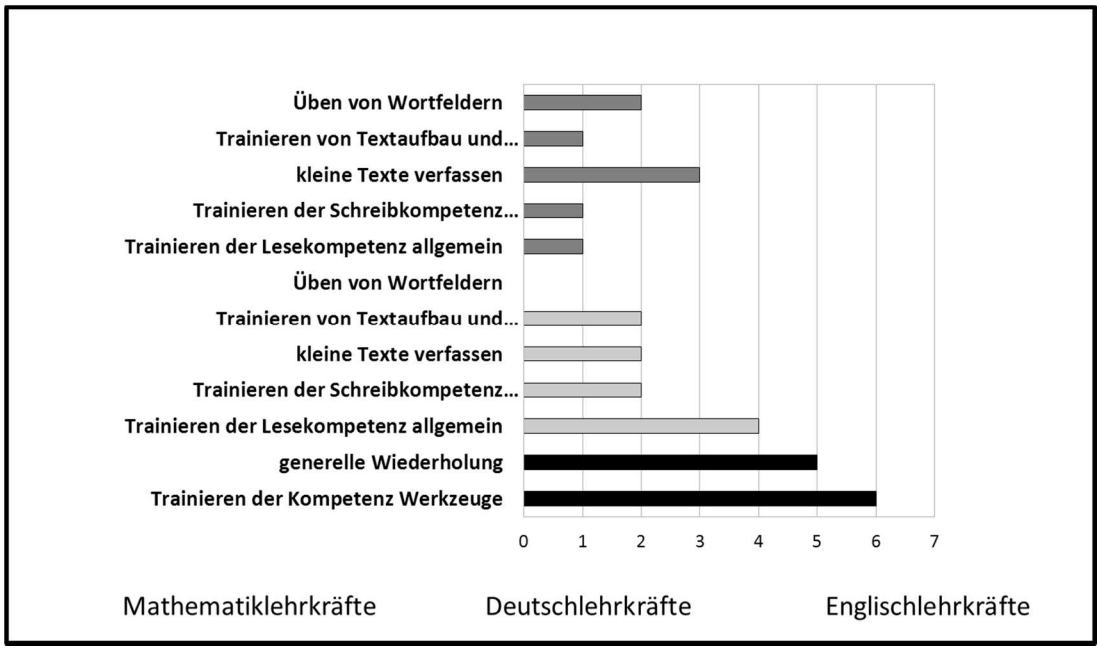


Abbildung 3.2: Kodierhäufigkeit in der Kategorie „Content Approach“

Der Umfang der Vorbereitung auf die Lernstandserhebungen im Jahr 2008 ließ sich leider nur in zwölf Fällen auf eine Stundenzahl abbilden (vgl. Abb. 3.3). Unter den zwölf Fällen sind auch die beiden Lehrkräfte, die gar kein Testcoaching durchführten. Die anderen zehn verteilen sich folgendermaßen: Fünfmal dauerte die Vorbereitung neun oder mehr Stunden. Viermal hatte die Vorbereitung einen Umfang von 6-9 Stunden, einmal nur 3-5 Stunden. Für alle zehn Fälle gilt aber, dass der Umfang der Vorbereitung über das hinausging, das man vielleicht allein mit notwendigen Erklärungen zum ungewohnten Verfahren begründen könnte. Für die sechs Fälle, zu denen keine Kodierungen zum Umfang gemacht werden konnten, lässt sich aufgrund der Zahl kodierter Maßnahmen (zwischen vier und acht) schließen, dass dies auch dort gegolten hat.

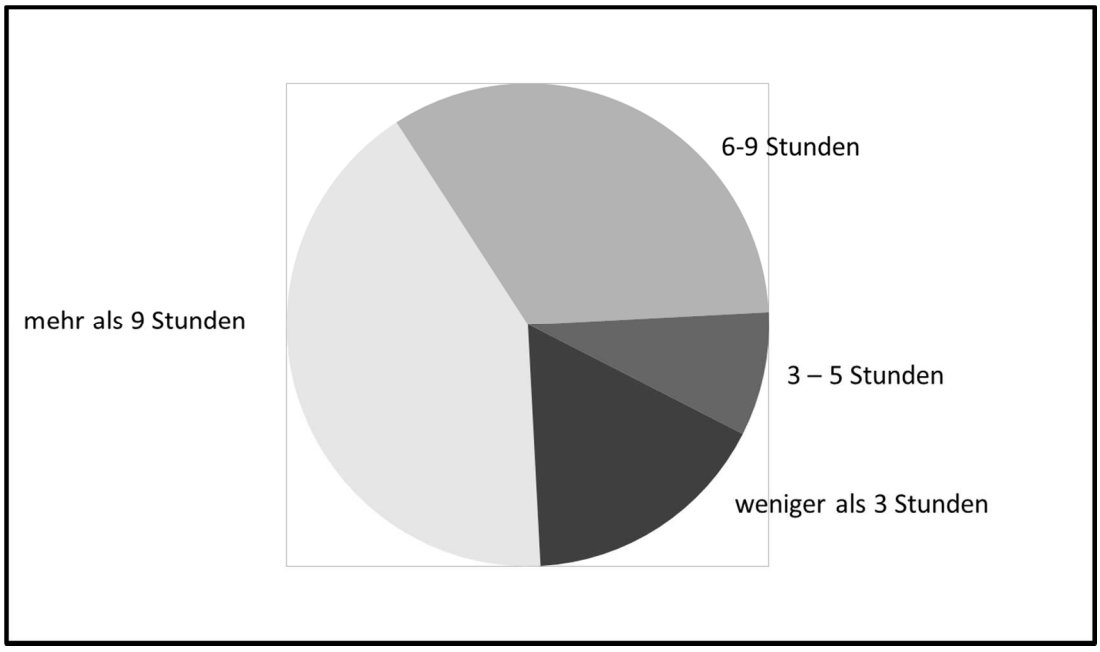


Abbildung 3.3: Ausprägungen in der Kategorie „Umfang“

Für dreizehn Protokolle zu Fachlehrerinterviews konnten Kodierungen in der Hauptkategorie „außerschulische Vorbereitung“ vorgenommen werden. Neun Lehrer waren der Meinung, ihre Schülerinnen und Schüler hätten weder besonders geübt noch speziell für die Lernstandserhebungen Nachhilfe in Anspruch genommen. Letzteres hatte von den achtzehn Lehrkräften nur eine beobachtet. Ihr Schüler und Schülerinnen haben nach ihrer Ansicht außerdem besonders zu Hause geübt. Die Kodierung „besonders intensives Üben“ konnte auch bei drei weiteren Protokollen vergeben werden.

Im Vergleich zu früheren Vorbereitungen gab eine Person an, ihre Vorbereitung sei im Jahr 2008 nicht anders verlaufen. Fünfmal wurden in Protokollen die Kodierung „Veränderung mit gleichem Umfang“ für die Vorbereitung vergeben. Drei Lehrkräfte bereiteten nach ihren Aussagen im Jahr 2008 umfangreicher vor. Eine der beiden Lehrkräfte, die keine Testcoaching-Maßnahmen durchführten, gab an, früher ihre Schüler und Schülerinnen vorbereitet zu haben, jetzt also in „geringerem Umfang“ vorzubereiten.

Von den sechzehn Lehrkräften, die Testcoaching durchführten, ordneten alle den Lernstandserhebungen mindesten eine der beiden Funktionen „Rückmeldung der Kompetenz an Schüler" und „Instrument zur Notenfindung" zu. In den Funktionen wird ausgedrückt, dass man die Ergebnisse der Lernstandserhebungen mit der Leistung der Schülerinnen und Schüler verbindet. Sie entsprechen der ersten Deutungsmöglichkeit von Prüfungsergebnissen als Rückmeldung an die Lernenden (Schrader & Helmke, 2002). Die vier Kodiermöglichkeiten „Rückmeldung über Unterricht zur weiteren Planung", „Arbeitszeugnis einzelner Lehrer", „Rückmeldung über Schulsystem" und „Ranking/Mittelzuweisung" könnte man grob als Ausdruck der Ergebnisse als eine Rückmeldung an die Lehrenden verstehen und gehören damit zur zweiten Deutungsart von

Prüfungsergebnissen. Diese vier Kodiermöglichkeiten konnten in zwölf Protokollen vergeben werden.

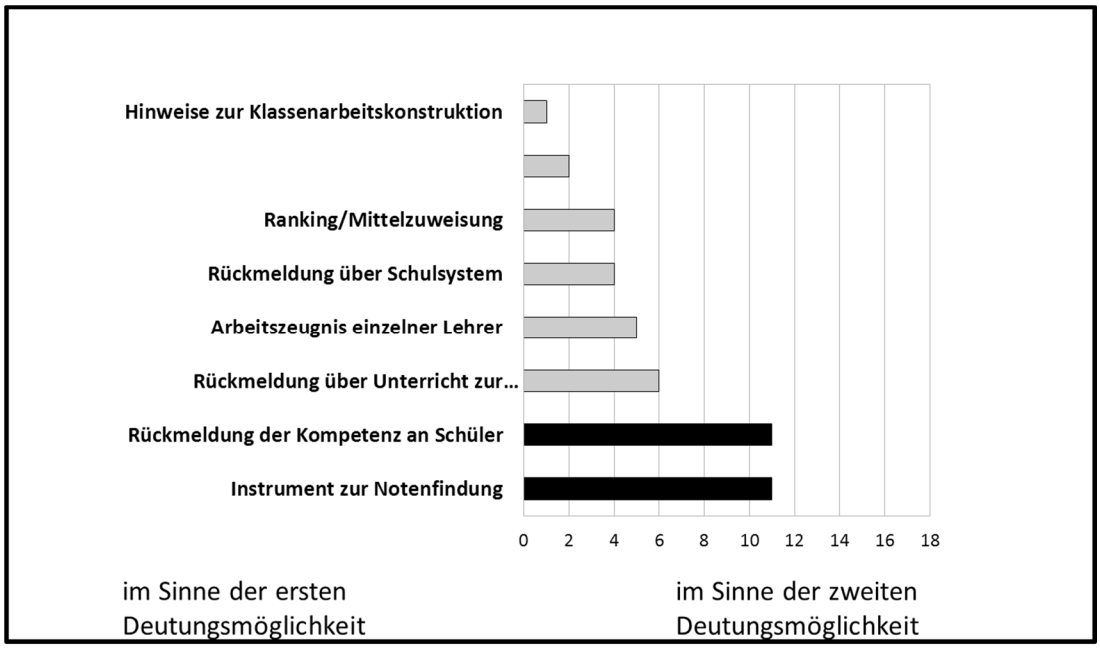


Abbildung 3.4 Kodierhäufigkeit in der Kategorie „Funktionen“

Insgesamt kann folglich festgehalten werden, dass die Lehrkräfte mehrheitlich die zur Vorbereitung auf die zentralen Lernstandserhebungen verfügbaren Möglichkeiten weitestgehend ausgenutzt haben. Die Vorbereitung fand dabei innerhalb der Unterrichtszeit statt, eine (zusätzliche) außerschulische Vorbereitung konnte nicht gefunden werden. Der durchschnittlich hohe Grad der Vorbereitung schien bei den befragten Lehrkräften dauerhaft zu sein. Das Vorbereitungsverhalten lässt sich als Reaktion auf die an Lehrkräfte gerichtete Rückmeldung über ihren Unterrichtserfolg deuten. Dafür spricht auch der hohe Stellenwert, den die befragten Lehrkräfte mehrheitlich den Lernstandserhebungen einräumten. Möglich ist aber auch eine Erklärung, die eine Vorbereitung als Maßnahme „im Sinne der Schülerinnen und Schüler“ annimmt. Nach dieser wird die Vorbereitung als Teil des Unterrichts gesehen.

4 Professionalität und Professionalisierung von Lehrkräften im Kontext von Unterrichten und Innovieren

Der dritte Blickwinkel dieser Arbeit neben zentralen Lernstandserhebungen als Steuerungsinstrument (Kap. 2) und Testcoaching als eine spezielle Unterrichtsqualität (Kap. 3) ist der auf die Lehrkraft und der Grad der individuellen Professionalisierung in Bezug auf die Kompetenzbereiche Unterrichten und Innovieren (vgl. die Standards für die Lehrerbildung der KMK für die Bildungswissenschaften, Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2004). Die Grundannahme an dieser Stelle ist die Bedeutung der Lehrerpersönlichkeit für das Handeln der Lehrkraft im Unterrichtskontext und damit mittelbar für den Lernprozess von Schülerinnen und Schülern mittels durch die Lehrkraft angebotener Lernumgebungen. Die in den sechziger bis achtziger Jahren auf dem Prozess-Produkt-Paradigma beruhenden Erklärungsmodelle (Haag, 2004) werden dabei in den Unterkapiteln (4.1) und (4.2) erweitert und mit dem Lehrer-Expertenparadigma (Bromme, 2008; Bromme & Haag, 2008) und dem Ansatz der Lehrer-Handlungskompetenz (Krauss et al., 2004) verbunden. Dies stellt die Verknüpfung der Lehrerpersönlichkeit mit dem Kapitel 3 dar. Differenzen im Testcoaching werden demzufolge unabhängig von der Komplexität des Instruments „zentrale Lernstandserhebungen“ als Teil der individuellen Unterrichtsgestaltung angesehen, die durch die der Lehrkraft individuell zur Verfügung stehenden Ressourcen gesteuert wird. Professionalität meint dabei den Umgang mit diesen für die Lehrtätigkeit immanenten Ressourcen.

Mit dem Unterkapitel (4.3) werden die zentralen Lernstandserhebungen als Steuerungsinstrument (Kap. 2) mit dem Unterrichtsgeschehen in Beziehung gesetzt. Zentrale Lernstandserhebungen werden hier als Feedbackangebot verstanden und die Vorbereitung auf die zentralen Lernstandserhebungen als Wirkung der (angekündigten oder bereits früher durchlaufenen) Feedbacksituation betrachtet. Die Professionalisierung[72] umfasst in diesem Sinne die Nutzung des Feedbackangebots als Teil der Innovationskompetenz, wie sie in den Standards der Lehrerbildung formuliert sind (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2004). Handlungen als Lösung gesellschaftlicher Probleme (hier der Unterricht) werden durch die Nutzung von Feedbackinformationen auf eine evidenzbasierte, rationale Basis gestellt (Berkemeyer & Bos, 2009). Die Innovationskompetenz kann ebenfalls durch eine Ressourcenkonzeption erklärt

[72] Mit dem hier verwendeten Begriff der Professionalisierung wird bereits unterstellt, dass es sich bei Lehrkräften um Professionelle handelt, wenn sie den in den Standards der KMK formulierten Aufgaben gerecht werden. Dies ist nicht als Antwort auf die Frage misszuverstehen, in welchem Grad der Berufsstand Lehrer eine Profession darstellt.

werden. Gleichzeitig kann aber auch eine Kopplung von Handlungsbereichen (hier: Feedbacknutzung und Testcoaching) analysiert werden.

Beide Zweige dieses Kapitels führen abschließend zu einem neuen Handlungskompetenzmodell, welches die Anforderungen an Lehrkräfte im Bereich Unterrichten und Innovieren integriert und für letzteres die besonderen Bestandteile herausstellt. Dieses Modell ist die Grundlage für die im zweiten Teil dieser Arbeit dargestellten Fragebogenstudien unter Mathematiklehrkräften.

4.1 Vom Prozess-Produkt-Paradigma zur Lehrer-Handlungskompetenz

Zu Beginn dieses Kapitels sollen die Forschungsparadigmen der Lehrerforschung kurz skizziert werden. Erstens handelt es sich um das Prozess-Produkt-Paradigma, welches das Persönlichkeitsparadigma abgelöst hat. Dieses bildet bis heute die Gruppe der vorherrschenden Untersuchungsdesigns der Lehrerforschung. Es wurde zweitens schließlich in den letzten zwanzig Jahren um das Expertenparadigma ergänzt. Der hiermit vorgeschaltete Einschub über die Forschungstraditionen hilft beim Verständnis, warum der empirische Schwerpunkt dieser Arbeit auf der Lehrerpersönlichkeit als erklärende Größe für die Qualität und die Quantität der Vorbereitung auf die zentralen Lernstandserhebungen liegt.

4.1.1 Prozess-Produkt-Paradigma

Das ursprüngliche Prozess-Produkt-Paradigma bezeichnet Erklärungsmodelle für den Unterrichtserfolg, in dem Lehrerhandlungen als Prozess bzw. unabhängige, empirisch zugängliche Variable und Schülerhandeln (vorwiegend der Lernerfolg, aber möglich ist auch die Interaktion zwischen Schülerinnen und Schülern) als Produkt bzw. abhängige und ebenfalls empirisch zugängliche Variable verstanden wurden (Haag, 2004). Unter der Fragestellung, welche Handlungsweisen von Lehrpersonen zum besten Unterricht (i.S. eines durch das Lösen von möglichst vielen Testaufgaben bewerteten Unterrichts) führen, wurden dabei einzelne Verhaltensweisen untersucht und gegen andere abgewogen, um daraus einfache Kausalbeziehungen abzuleiten. Das Prozess-Produkt-Paradigma löste das Paradigma der Lehrerpersönlichkeit aus der Mitte des zwanzigsten Jahrhunderts ab. Jenes fragte nach den positiven Persönlichkeitseigenschaften, war darin aber wenig theoriegeleitet und methodisch unreflektiert (Bromme, 1997).

Nach Haag hat sich mittlerweile allerdings auch die Einsicht durchgesetzt, dass die „universal gute Lehrkraft“ nicht durch Variablen dieser Art zu identifizieren ist, wie sie nach dem Prozess-Produkt-Paradigma untersucht wurden. Unterschiede im Lehrerhandeln und im Lehrererfolg sind nicht allein über Beobachtungen einzelner Handlungselemente von Lehrkräften möglich. Die konkreten Schulkontextbedingungen erfordern ein hohes Maß an Adaptivität. Die Lehrkraft kann zwar als zentrale Erklärungsgröße für gelungenen Unterricht angesehen werden, neben dem Verhalten müssen aber auch ihre Überzeugungen (u.a. Wissen und Können) berücksichtigt werden, da während des Lehrer-Handelns die Lehrperson situationsabhängig aus verschiedenen Handlungsalternativen wählen können muss. Lehrerhandeln und -entscheidungen werden außerdem durch das Denken der Lehrkraft vor, bei und nach der Unterrichtshandlung beeinflusst. Auf die Ineffektivität des Versuchs, einzelne Variabeln für guten Unterricht zu finden, haben insbesondere auch Lehrkräfte hingewiesen, die den Ertrag dieser Forschungsrichtung als nicht hilfreich erkannten (Berliner, 1990; Bromme, 1997, 2008; Haag, 2004; Hugener, 2008; Verloop, Driel & Meijer, 2001).

Neben der behavioristischen Sichtweise und dem Anspruch der Generalisierbarkeit und Allgemeinheit wurde erkannt, wie stark der Lernerfolg von den jeweiligen Schülerinnen und Schülern abhängig ist (Hugener, 2008). Zwar beeinflusst die Lehrkraft durch sein Handeln das unmittelbar sichtbare Verhalten der Schüler und Schülerinnen, ein beabsichtigter Lernprozess ist durch das sichtbare Verhalten aber noch lange nicht garantiert. In Folge dessen sieht man die Lehrkraft als für Lerngelegenheiten verantwortlich, nicht mehr für den tatsächlichen Lernerfolg der Schüler und Schülerinnen. Wenn man den individuellen Informationsverarbeitungsprozess von Schülerinnen und Schülern berücksichtigen möchte, kann das Prozess-Produkt-Modell als Prozess-Meditations-Produkt-Modell aufgefasst werden. Das Angebot-Nutzungs-Modell von Helmke (s.a. Helmke, 2009) ist ein Prototyp dieser Modellklasse, in der besonders die aktive Lernzeit und die Tiefe der Auseinandersetzung mit dem Unterrichtsgegenstand als wichtiger Aspekt betrachtet werden (Brunner, Kunter, Krauss & Klusmann et al., 2006). Das Lehrerhandeln ist aber zwei strukturellen Unsicherheiten ausgesetzt. Neben der Unsicherheit ob des Lernerfolgs der Schülerinnen und Schüler ist auch das Schaffen von Lerngelegenheiten von der Kooperation der Schüler und Schülerinnen abhängig. Aufgrund der doppelten strukturellen Unsicherheit wäre ein Opportunitäts-Nutzungsmodell mit doppelter Kontingenz (Wirkung der Schülerinnen und Schüler auf Unterrichtsangebot und -nutzung) angebracht (Baumert & Kunter, 2006; Berliner, 1990).

Mit der Erweiterung des Prozess-Produkt-Modells wechselte die Perspektive von einer kollektiv-systemischen Sicht zu einer individualpsychologischen Sichtweise, während die Person der Lehrkraft im Prozess-Produkt-Paradigma in ein Bündel von Teilfertigkeiten zerfiel, wie Bromme es nennt (Bromme, 1997; Brunner, Kunter, Krauss & Klusmann et al., 2006). Eine andere Modifikation des Prozess-Produkt-Schemas ist ein höheres Abstraktionsniveau

bei der Unterrichtsbeobachtung. Oser und Baeriswyl unterscheiden begrifflich zwischen Sichtstruktur und Tiefenstruktur als Ebenen der Unterrichtsbeschreibung (Oser & Baeriswyl, 2001). Die Sichtstruktur umfasst dabei die direkt durch die Organisationsstruktur gegebenen Kategorien wie methodische Unterrichtselemente. Unter Tiefenstruktur werden übergreifende Erfolgsziele verstanden, die als Bestandteile einer lernförderlichen Unterrichtsatmosphäre identifiziert werden konnten (Kunter et al., 2006). Dazu zählen effektive Klassenführung, ein zielorientierter Unterricht, ein an die Verarbeitungsgeschwindigkeit der Schüler und Schülerinnen angepasstes Unterrichtstempo und kognitiv herausfordernde Aufgaben, aber auch Anerkennung, Motivation und Interesse fördernde Unterstützung (Brophy, 1999; Helmke, 2009).

4.1.2 Lehrer-Expertenansatz

Die individualpsychologische Sichtweise ist vor allem eine kognitive. Der Lehrer-Expertenansatz lässt sich unter die Kognitionsforschung subsumieren. Die Forschungsrichtung kann als Lehrerkognitionsforschung beschrieben werden. Nach Dann werden folgende Grundannahmen getroffen: a) Lehrer sind autonom und verantwortlich Handelnde sowohl in der Erfüllung der berufsalltäglichen Aufgaben als auch in der Fortentwicklung ihrer persönlichen Praxis, b) Lehrkräfte verfolgen bei ihrem Handeln Ziele für ihre Klienten (Schülerinnen und Schüler, Schulleitung, Erziehungsberechtigte), c) beim zielgerichteten Handeln ist ihr Handlungsspielraum aktiv-kognitiv strukturiert, d) Lehrkräfte - im Sinne von Lehrern als Experten - greifen dabei auf professionelles Wissen zurück und e) ist dieses Wissen teilweise kohärent zu gesellschaftlich geteiltem Wissen (Dann, 2008).

Dementsprechend unterstellt der Lehrer-Expertenansatz Lehrkräften ein (Unterrichts-) Handeln, welches auf Wissen und Können beruht, und spricht von der Profession des Lehrens und Lernen. Die berufliche Erfahrung der Lehrkräfte wird dem wissenschaftlichen Wissen nicht gegenüber gestellt, sondern als Transformator zu jedem Professionswissen verstanden (Bromme, 2008). Der Lehrer-Expertenansatz erweitert das vorher existierende Prozess-Produkt-Paradigma um die Sichtweise eines professionellen, problemorientierten Zugangs zum Unterrichten. Es fragt nach den Wegen, berufsbedingte Aufgabenstellungen zu bewältigen und sich das dafür nötige Wissen und Können anzueignen. Typische Untersuchungsgegenstände sind die Konzepte (oder auch: Unterrichtsskripts) von Lehrkräften zu bestimmten Unterrichtssituationen. Diese werden beispielsweise durch das Vorspielen von Unterrichtssequenzen erfasst, die von Lehrkräften fortgesetzt und kommentiert werden sollen.

Zusammen mit den Unterrichtsskripts stehen aber auch komplexe Interaktionsmuster und zeitlich und sachlich überdauernden Bedeutungsstrukturen als Untersuchungsgegenstand im Blickfeld. Statt der einzelnen Handlungen einer Unterrichtsstunde rücken die Unterrichtsabläufe und die elaborierten Ziele in den Vordergrund, welche mehr als eine

einzelne Unterrichtsstunde überdauern. Dieses wird speziell als Anwendung von Expertenwissen verstanden. Drittens werden mit dem Expertenansatz konstruktivistische Vorstellungen integrierbar. So genannte „subjektive Theorien“ (s. (4.2.2.1)) bzgl. ihrer Wirkungen auf das Lehrerverhalten können untersucht werden (Bromme, 1997).

Neben dem Ansatz, Lehrkräften generell Expertise im Bereich des Lehrens und Lernens zuzusprechen, kann Experte auch der Spitzenkönner sein (Bromme, 2008). Dieser Ansatz sieht Lehrkräfte nicht zu Erziehungsberechtigten bzw. Schülerinnen und Schülern in einer Experten-Laien-Relation, sondern unterscheidet innerhalb der Menschen mit Expertise zwischen Novizen, die sich am Anfang ihrer Entwicklung befinden, und Experten, die den Entwicklungsprozess bereits durchlaufen haben. Beide Ansätze haben aber dasselbe Ziel: die gute Lehrkraft zu identifizieren (Besser & Krauss, 2009).

Problematisch am Lehrer-Expertenansatz ist die Frage, wann eine Lehrkraft als Experte zu betrachten ist. Die Einordnung unter dem Adaptivitätsaspekt kann ebenfalls nicht generalisiert werden und ist für jede Untersuchung neu zu treffen. Das Ausmaß der beruflichen Erfahrung allein zeigte sich dabei als eher ungünstiger Indikator im Vergleich zur Schülerleistung (Brunner et al., 2006), wenngleich der Umfang der Berufserfahrung das Ausmaß der Lerngelegenheiten widerspiegelt und somit ein gewisser Zusammenhang der Berufsjahre mit der Schülerleistung besteht. Mit Blick auf die oben dargestellte Problematik bzgl. der Testleistung und des Testcoachings ist aber auch dieses Kriterium kritisch zu sehen. Auch der Erfolg von Schülerinnen und Schülern in Tests ist situationsabhängig und in jedem Fall eine indirekte Messung. Zusätzlich wirkt sich in Modellen, die die Testleistung von Schülerinnen und Schülern als Kriterium für die Expertise von Lehrkräften nehmen, die von Baumert und Kunter angesprochene doppelte Unsicherheit (s. (4.1.1) aus. Folglich ist eine „gute Lehrkraft“ nicht zwingend eine „erfolgreiche Lehrkraft“ (Berliner, 2001). Ein weiteres Problem ist die Konzentration auf begünstigende, erwerbbare Faktoren. Bromme weist daraufhin, wie wichtig auch die Erforschung von hinderlichen Faktoren ist (Bromme, 2008). Zu diesen können beispielsweise die erlebte berufliche Belastung oder die unter dem Begriff Big-Five zusammengefassten Persönlichkeitsmerkmale gehören.

Der Lehrer-Expertenansatz kann „Experte“ noch in einem anderen Sinne betrachten, nämlich als Experte für das eigene Handeln, die eigenen Absichten, Überzeugungen und Emotionen. Dann sieht in Lehrkräften folglich potenzielle Nutzer von Wissen und Technologien der Unterrichtsforschung, die sich mit deren Angeboten aktiv und konstruktiv auseinandersetzen und im Austausch und teilweise in Unterrichtsforschung eingebunden zur Entwicklung des professionellen Wissensfundus beitragen (Dann, 2008).

4.1.3 Lehrer-Handlungskompetenz

Wenn Lehrkräften zielgerichtetes Handeln mit Rückgriff auf ihr Können und Wissen (Kompetenz im engeren Sinne nach Weinert) unterstellt wird (Bromme, 2008; Dann, 2008), impliziert dies notwendigerweise ein Handlungskompetenzmodell. Die Modelle müssen aber mindestens auch Überzeugungen (als Werthaltungen, subjektive Theorien und Ziele) sowie motivationale Orientierungen und selbstregulative Fähigkeiten (das Berufserleben und Kompetenz- und Kontrollüberzeugungen) berücksichtigen. Dadurch wird das Professionswissen zu einem „Kompetenzverständnis im weiteren Sinne“ ergänzt (Weinert, 2001).

Ein solches Modell muss einerseits erklären, in welchem Verhältnis Wissen und Können zu den Zielen und den subjektiven Erfahrungen stehen, es muss gleichzeitig darstellen, in welcher Weise diese auf das Lehrerhandeln wirken. Vor allem muss es aber die Einflussfaktoren systematisieren. Diese Systematisierung gestaltet sich allerdings als äußerst schwierig, da die verschiedenen Einflussgrößen nicht eindeutig separiert werden können und diese auch uneinheitlich bezeichnet werden.

Das erweiterte Handlungskompetenzmodell dieser Arbeit basiert auf verschiedenen Studien, die in den letzten zehn Jahren zur Lehrerprofessionalität im deutschen Sprachraum durchgeführt wurden. COACTIV[73] untersucht den Zusammenhang von Professionswissen der Lehrkräfte, Überzeugungen über das Lehren und Lernen und die Genese von Mathematik und die Selbstregulationsfähigkeit der Lehrkräfte zu ihrem Unterrichtshandeln bzw. der Wahrnehmung des Unterrichtsstils durch die Schüler und Schülerinnen (Dubberke, Kunter, McElvany, Brunner & Baumert, 2008; Klusmann, 2008a; Krauss et al., 2004). MT21[74] bzw. TEDS-M untersuchen bei in der Ausbildung befindlichen Lehrkräften ebenfalls das Professionswissen, Überzeugungen über das Lehren und Lernen, Überzeugungen zu den Zielen des Mathematikunterrichts, Kontroll- und Kompetenzüberzeugungen und die Berufswahlmotivation (Blömeke, Kaiser & Lehmann, 2008, 2010a). Überzeugungen über das Lehren und Lernen und die Genese von Mathematik sowie die Selbstregulationsfähigkeit und die motivationale Orientierung waren auch Gegenstand verschiedener anderer Untersuchungen (beispielsweise: Diedrich, Thußbas & Klieme, 2002; Schaarschmidt, 2009; Schmitz, 2000). Daraus ergeben sich fünf Bereiche, von denen die ersten vier im Handlungskompetenzmodell von COACTIV (Baumert & Kunter, 2006) aufgegriffen wurden: (1) Wissen und Können, (2) Überzeugungen über Ziele, Werte und Kausalbeziehungen, (3)

[73] Professionelle Kompetenz von Lehrkräften, kognitiv aktivierender Unterricht und die Kompetenz von Schülerinnen und Schülern (COACTIVE): eine Studie des Max-Plank-Instituts für Bildungsforschung in Kooperation mit der Universität Kassel und der Universität Oldenburg.

[74] Teacher Education and Development Study in Mathematics (TEDS-M) ist ein internationals Forschungsprojekt. Die Verantwortung in Deutschland obliegt (Sigrid Blömeke und Rainer Lehmann (Humboldt-Universität zu Berlin) sowie Gabriele Kaiser (Universität Hamburg). Mathematics Teaching in the 21st Century (MT21) ist die zugehörige Pilotstudie.

motivationale Orientierung, (4) Regulationsfähigkeit und (5) Innovationsfähigkeit (ergänzte Auflistung).

Zu diesen ursprünglich vier Bereichen sowie zum ergänzten fünften Bereich, der Innovationsfähigkeit, werden im Folgenden jeweils der theoretische Rahmen und ausgewählte Forschungsergebnisse vorgestellt. Die Darstellungen der Bereiche „Wissen und Können“ sowie „Überzeugungen“ sind dort genauso ausführlich dargestellt wie die anderen drei Bereiche, da ihnen grundsätzlich für das Unterrichtshandeln der Lehrkräfte große Bedeutung zukommt, sie spielen aber in den Studien dieser Arbeit nur eine untergeordnete Rolle.

4.2 Facetten der Lehrerpersönlichkeit und ihre Auswirkungen auf den Unterricht

Im zweiten Teil dieses Kapitels werden nacheinander die vier Bereiche (1) Wissen und Können, (2) Kausal- und Zielüberzeugungen, (3) berufliches Erleben und (4) Kompetenz- und Kontrollüberzeugungen erläutert (vgl. auch Abb. 4.1). Ausgangspunkt der Betrachtung ist dabei jeweils erst einmal das für COACTIV aufgestellte Modell der Lehrer-Handlungskompetenz von Krause u.a. (2004). Im Abschnitt zu Wissen und Können wird zuerst erläutert, wie Wissen und Können angeeignet werden, dann werden verschiedene Domänen differenziert. Im Abschnitt zu Kausal- und Zielüberzeugungen werden zuerst verschiedene Begriffe (Überzeugungen, Einstellungen, subjektive Theorien) zueinander gegenüber gestellt, um anschließend die Wirkungsweise auf das Lehrerhandeln im Unterricht zu erläutern. Beide Bereiche werden zusammen als gegenstandsbezogene Überzeugungen aufgefasst. [75] Anschließend folgen die personenbezogenen Überzeugungen. Hier wird zuerst die Einteilung des COACTIV-Modells von motivationaler Orientierung und Selbstregulation neu geordnet. Unterschieden wird zwischen Überzeugungen des beruflichen Erlebens, die die drei Dimensionen Arbeitsengagement/Arbeitszufriedenheit, Widerstandsfähigkeit und berufliche Emotionen umfassen, und den Kompetenz- und Kontrollüberzeugungen, die das Fähigkeitsselbstkonzept und die Selbstwirksamkeitserwartung vereinen. Beide Überzeugungsklassen bilden ein Ressourcenbündel, das als notwendige Voraussetzung für „guten Unterricht“, also auch für eine gute Vorbereitung auf zentrale Lernstandserhebungen angesehen wird. Das theoretische Modell dazu ist das Job-Demand-Resource-Modell von Schaufeli und Bakker (2004).

[75] Zusammen mit den in Abschnitt (4.2.2) dargestellten Überzeugungen werden Wissen und Können später ebenfalls als gegenstandsbezogene Überzeugungen aufgefasst. Für die Nutzung des Begriffs „Überzeugungen“ auch für diese Art von Kognitionen spricht die individuelle Repräsentation (Sembill & Seifried, 2009).

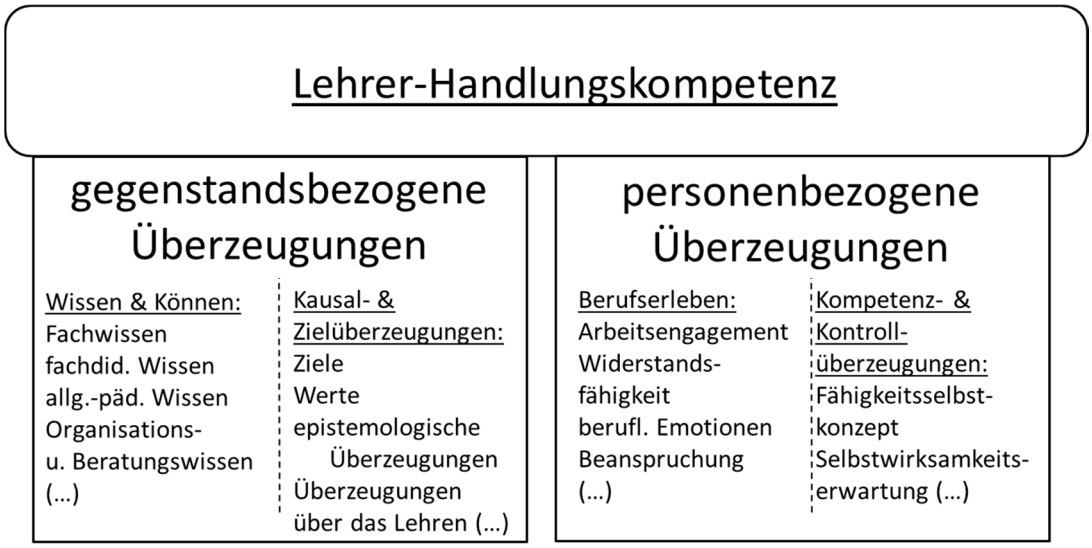


Abbildung 4.1 Ausgangsmodell der Lehrer-Handlungskompetenz

4.2.1 Wissen und Können (Fachwissen, fachdidaktisches Wissen, pädagogisches Wissen, Organisationswissen oder auch Interaktionswissen, Beratungswissen)

Das Expertenwissen bzw. Wissen und Können von Lehrkräften können qualitativ unterschieden und in einzelne Domänen eingeteilt werden. Verloop, Driel und Meijer differenzieren qualitativ und sprechen einerseits von „kowledge for teachers“ als dem Wissen, welches Lehrkräften durch die Wissenschaft und Ausbilder zur Verfügung gestellt wird, und andererseits von „knowledge of teachers“ als dem Wissen, welches aus dem Knowledge for Teachers durch eigene Erfahrungen, die eigenen Persönlichkeit und den eigenen Lernprozess entsteht (Verloop et al., 2001). Das Knowledge of Teachers hat eine lose Ähnlichkeit mit so genannten „subjektiven Theorie“ (s.u.), weil es gleichsam konstruiertes Kausalzusammenhänge beschreibt, und beruht (aber) auf wissenschaftlicher Forschung.

Verloop u.a. listen eine Reihe von Studien auf, die die verschiedenen Sichtweisen auf das Wissen und Können von Lehrkräften offenbaren. Wissen kann dem folgend als “personal knowledge” “professional craft knowledge”, “action oriented knowledge” und “content and context related knowledge” bezeichnet werden, wodurch ausgedrückt wird, dass es individuell ist, in der Praxis erworben wurde, aber auch für diese relevant und situationsabhängig ist (Verloop et al., 2001).

Nach Baumert und Kunter zeichnet sich die Struktur von Expertenwissen nach aktuellem Forschungsstand durch fünf Charakteristika aus: (1) professionelles Wissen ist

domänenspezifisch und ausbildungs- bzw. traningsabhängig, (2) Expertenwissen ist sehr gut vernetzt und hierarchisch organisiert, (3) in professionellen Domänen ist Expertenwissen um Schlüsselkonzepte und eine begrenzte Zahl von Ereignisschemata arrangiert, an die Einzelfälle, episodische Einheiten oder Sequenzen von Episoden (Skripts) angedockt sind, (4) professionelles Expertenwissen integriert Kontexte und erlaubt variantenreicheres „opportunistisches Verhalten", (5) Basisprozeduren sind automatisiert, aber gleichwohl flexibel an die spezifischen Bedingungen des Einzelfalles und des Kontextes adaptierbar (Hatano, Inagaki, 1986, nach Baumert & Kunter, 2006).

Die Genese von Professionswissen

Professionsspezifisches Wissen wird nach und nach im Zuge der professionellen Entwicklung aufgebaut und vertieft. Für Hascher und Krapp zeichnet sich der Entwicklungsprozess der Lehrkräfte durch permanente Veränderung aufgrund selbstgesteuerten Lernens aus. Selbstgesteuertes Lernen sei notwendig, da für Lehrkräfte keine systematischen „von oben" gesteuerten Maßnahmen der Personalentwicklung verfügbar seien. Daher sei die Entwicklung der Lehrkräfte stark von ihrer Eigeninitiative abhängig (Glaser, 1996; Hascher & Krapp, 2009).

Dies entspricht dem Genesemodell von Bereiter und Scardamalia (1993). Der Ansatz sieht vor, dass Professionswissen gewonnen wird, indem kognitive Ressourcen in das Lösen von Problemen investiert werden. Dies geschieht stetig und mit einer wachsenden Komplexität, sodass mit der Zeit Problemlöseprozesse automatisiert werden und Ressourcen zur Lösung komplexerer Probleme bereitstehen. Entscheidend in diesem Ansatz ist die gezielte, bewusste Suche nach komplexeren Aufgaben. Eine Expertemdomäne zeichnet sich durch ihre unendlich komplexen Fundamentalprobleme aus. Für das Unterrichten ist dies beispielsweise die Beseitigung von Unwissenheit („elimination of ignorance") (Bereiter & Scardamalia, 1993). Dies führt begrifflich zu der Unterscheidung von „Laien" und „Experten". Experten-Novizen-Unterscheidungen werden von Bereiter und Scardamalia explizit abgelehnt, denn diese vermittelten den Eindruck, beide arbeiteten an der Lösung derselben Probleme (Bereiter & Scardamalia, 1993). Dies trifft aber nur auf die Fundamentalprobleme zu. Im Sinne des dynamischen Verständnisses steht aber auch der Experte innerhalb der Fundamentalprobleme vor immer neuen Herausforderungen und Problemen statt ausschließlich vor Aufgaben, die lediglich einen Novizen herausfordern.

Der zweite Ansatz stammt von Ericsson und Charness sowie Ericsson, Krampe und Tesch-Römer (1994 bzw. 1993). Tonangebend ist hier die Vorstellung einer Wandlung des Novizen zum Experten. Ihr Ansatz beruht auf zeitlich sehr intensiven Lern- und Übungsaktivitäten (so genannten „deliberate practice"). Diese Lern- und Übungsgelegenheiten sind auf stetig wachsende Leistung ausgerichtet und erfordern große persönliche Anstrengung. Die zugrunde liegende Motivation ist folglich eher einem Leistungsmotiv ähnlich und weniger intrinsisch zu erklären. Gleichzeitig ist der ganze Lernprozess sehr langfristiger Natur und es

dauert mindestens zehn Jahre, bis ein merkbares Expertise-Level erreicht ist. Ericsson, Krampe und Tesch-Römer stellen die Bedeutung von (zeitnahem) Feedback zur Leistung für einen gelungenen Lernfortschritt heraus. Fehlt es an adäquatem Feedback, ist ihrer Ansicht nach kein gelungener Lernprozess möglich und Lernzuwächse sind nur bei äußerst motivierten Lernenden zu erwarten (Ericsson & Charness, 1994; Ericsson, Krampe & Tesch-Römer, 1993).

Ericsson und Kollegen weisen allerdings auf zwei Schwierigkeiten hin, die das Lernen durch Übung bedingen. Übungsphasen bedeuten das Risiko, Fehler zu begehen und zeitweise schlechtere Leistungen zu erbringen, und sie erfordern Investitionen (Ericsson et al., 1993). Es ist daher auch aus Sicht der Lehrkräfte rational, bewährte Handlungsmuster zu verwenden. Neue Unterrichtsstile oder auch nur neue Aufgabenarten auszuprobieren, setzen gravierende Gründe voraus, wenn eine Lehrkraft einmal ein gewisses Level an Expertise erreicht hat. Anderenfalls ist der mögliche Schaden für Schülerinnen und Schüler und Lehrkräfte zu groß.

Brunner und Kollegen fassen zusammen: "[B]eiden Ansätzen [ist] die Annahme gemein, dass Erfahrung in einer Domäne alleine nicht ausreicht, um domänenspezifisches Professionswissen zu erwerben. Entscheidend ist stattdessen, dass kognitive und motivationale Ressourcen gezielt in das Lösen von domänspezifischen Problemen beziehungsweise wohl strukturierten Übungsaktivitäten in der jeweiligen Domäne investiert werden." (Brunner et al., 2006 , S. 526).[76]

Als Beispiel nennen Brunner und Kollegen die Reflexion der Unterrichtsskripte im Kollegium wie es in Lesson Studies unter japanischen Lehrkräften geschieht. Dazu ist Wissen notwendig, um die Analyse bewältigen, die Ergebnisse darstellen und diskutieren zu können und um Schlussfolgerungen aus den Ergebnissen ziehen zu können (Brunner et al., 2006).

Unabhängig davon, ob man der Experten-Laien- oder der Experten-Novizen-Vorstellung anhängt, sind systematische Lernprozesse von Nöten. Kognitive und motivationale Ressourcen zu investieren, setzt Wissen voraus, wie Erfahrung sinnvoll genutzt werden kann. Anderenfalls führt der Versuch, aus der Praxis zu lernen, zu keinem Erfolg und wirkt demotivierend. Im schlimmsten Fall werden die falschen Folgerungen gezogen und negative Entwicklungen folgen. Schelten nennt dem folgend die Innovationskompetenz als eine Wissens- und Könnensdomäne neben dem fachlichen und dem didaktischen Wissen und Können. Mit Innovationskompetenz verknüpft er die Entwicklung neuer Bildungsinhalte und Bildungsformen sowie die Bestimmung neuer Bildungsziele (Schelten, 2009).

Das von Glaser als eine der zentralen Fähigkeiten bezeichnete „structure knowledge“ kann als ein Teil davon verstanden werden (1996). Die Fähigkeit, sein eigenes Wissen zu strukturieren und zu erweitern gehört damit genauso zu einer Kompetenz wie das inhaltliche Wissen. Dies ist nicht nur als Aufgabe für das Unterrichten von Schülerinnen und Schülern zu sehen, sondern gilt gleichermaßen für den (zukünftigen) Experten selbst. Nach Berliner

[76] Gleichwohl steigt mit der Berufserfahrung die Wahrscheinlichkeit, Professionswissen zu erbwerben.

zeichnet sich die Entwicklung zum Experten durch die Fähigkeit, Handlungsprozesse von einer konkreten Situation auf eine andere zu transferieren, durch eine stetig wachsende Menge von automatisch ablaufenden Handlungen und durch eine größere Flexibilität aus (Berliner, 2001).[77]

Wenn im nächsten Abschnitt die Wissensstrukturen dargestellt werden, wird deutlich werden, dass Wissen und Können zur Weiterentwicklung der eigenen Expertise in keinem der dargestellten Wissensmodelle eine explizite Position einnehmen. Obwohl die Profession „Lehrkraft" die Weiterentwicklung inkludiert und Innovation Chance wie Belastung sein kann (Altrichter, 2000; Kiper, 2009), werden Lehrkräfte stets als situativ geforderte Professionelle verstanden, deren Aufgabenfeld neben dem Unterrichten mit Eltern- und Schülergesprächen ausschließlich diejenigen Bereiche zu umfassen scheint, in denen Lehrkräfte gegenüber anderen einen Wissensvorsprung zu besitzen scheinen. Die im erweiterten Handlungskompetenzmodell (4.4) ergänzte Innovationskompetenz integriert diese Überlegungen zur Entwicklung von Expertenwissen.

Domänen des Professionswissens von Lehrkräften

Die ursprüngliche Differenzierung von Shulman, auf die beispielsweise das COACTIV-Modell rekurriert, sieht sieben Wissensdomänen des Lehrers vor. Neben dem Wissen über den Unterrichtsgegenstand (Content Knowledge) , dem fachdidaktischen Wissen (pedagogical Content Knowledge) und allgemeinem pädagogischen Wissen (general pedagogical Knowledge) unterscheidet Shulman noch das Wissen über das Fachcurriculum (Curriculum Knowledge), das Wissen über das Lernen allgemein (Knowledge of Learners), das Organisationswissen (Knowledge of educational Context) und das Wissen über Erziehungsziele, bildungstheoretisches und bildungshistorisches Wissen (Shulman, 1986)), tatsächlich bilden aber nur die ersten drei Komponenten einen Konsens in den verschiedenen aufbauenden Modellen (Baumert & Kunter, 2006).

Für COACTIV (Krauss et al., 2004) und TEDS-M (Blömeke, Kaiser & Lehmann, 2010b) wurden Wissens-Modelle der Lehrer-Handlungskompetenz entwickelt. Sie beruhen auf der Vorstellung, dass Wissen und Können (also deklaratives, prozeduales und strategisches Wissen) eine zentrale Komponente der Handlungskompetenz von Lehrkräften darstellen. Unterricht zu erteilen, steht darin als Kernaufgabe von Lehrkräften im Zentrum der Überlegungen (Baumert & Kunter, 2006). Beide Modelle nutzen die von Shulman (1987) vorgenommenen Unterscheidung des Professionswissens (Fachwissen, fachdidaktisches Wissen und pädagogisches Wissen), das Modell für COACTIV ergänzt diesen Wissenskanon

[77] Er wählt dazu den Vergleich mit einem Profi-Golfspieler. Hierin wird allerdings ein deutlicher Unterschied zu Lehrkräften als Experten für das Lehren und Lernen deutlich. Die Weiterentwicklung eines Profi-Sportlers geschieht eigentlich ausschließlich mittels Trainer, welches für Lehrkräfte nur eine Ausnahme darstellt. Lehrkräfte vollziehen nach Abschluss des Referendariats ihre Weiterentwicklung in der Regel nur während interner und externer Fortbildungen angeleitet, vorwiegend aber selbstständig.

zusätzlich um Organisationswissen (Fried, 2002; Shulman, 1987) und Beratungswissen (Bromme & Rambow, 2001; Krauss et al., 2004; Rambow & Bromme, 2000).

Fachwissen und fachdidaktisches Wissen

Die Bedeutung des Fachwissens umfasst u.a. einen Überblick über die Struktur des Fachs, um Lernziele setzen und Unterrichtsinhalte gewichten zu können, aber auch um Lernen am Vorbild zu ermöglichen. Es beinhaltet die Elemente des Schulstoffs und fachliches Alltagswissen, reicht aber darüber hinaus. Fachdidaktisches Wissen ist das Wissen, welches es ermöglicht, das Fachwissen anderen verständlich zu machen (Shulman, 1986). Es beinhaltet den Verhandlungs- und Vermittlungsaspekt (Wissen über adäquate Erklärung und Repräsentation fachlicher Inhalte), den Inhaltsaspekt (Wissen über das Potenzial des Schulstoffs für Lernprozesse) und den Schüleraspekt (Wissen über typische Schülerfehler und Schülerschwierigkeiten) (Brunner, Kunter, Krauss & Klusmann et al., 2006). Problemorientiert können beispielsweise drei Komplexe für fachdidaktisches Wissen benannt werden: (1) Wissen über didaktisches und diagnostisches Potenzial von Aufgaben, Wissen über die kognitiven Anforderungen und impliziten Wissensvoraussetzungen von Aufgaben, ihre didaktische Sequenzierung und langfristige curriculare Anordnung von Stoffen, (2) Wissen über Schülervorstellungen (Fehlkonzeptionen, typische Fehler, Strategien) und Diagnostik von Schülerwissen und Verständnisprozessen und (3) Wissen über multiple Repräsentations- und Erklärungsmöglichkeiten (Baumert & Kunter, 2006).

Während Baumert und Kunter herausstellen, dass die fachlichen Kenntnisse das fachdidaktische Wissen positiv beeinflussen können und gar notwendige Bedingung für qualitätsvollen Unterricht sei, ergaben die Untersuchungen in COACTIV und TEDS-M teilweise differenziertere Befunde zum gemeinsamen Auftreten von Fachwissen und fachdidaktischem Wissen von Mathematik-Lehrkräften. Zwar korreliert Fachwissen und fachdidaktisches Wissen teilweise stark (bei angehenden deutschen Lehrkräften r.70) (Blömeke et al., 2010), einige Realschulehrkräfte konnten aber in den Studien auch ein hohes fachdidaktisches Wissen nachweisen, ohne gleichzeitig über ein entsprechendes Fachwissen zu verfügen (Blum, Krauss & Neubrand, 2008), und in einigen Staaten war dieser Zusammenhang bei TEDS-M deutlich schwächer als für angehende deutsche Lehrkräfte (z.B. Schweiz r.40, Botswana r.18) (Blömeke et al., 2010). Döhrmann, Kaiser und Blömeke sehen die wissenschaftliche Diskussion über die Abgrenzung folglich auch noch nicht abgeschlossen. Es kann immer noch nicht mit großer Sicherheit belegt werden, ob fachwissenschaftliche Wissen und fachdidaktisches Wissen überhaupt separiert werden können. Sie sehen aber ähnlich wie Baumert und Kunter in einem Vergleich der TEDS-M-Ergebnisse mit Ergebnissen der so genannten Population II aus TIMSS2007 Anzeichen für einen Zusammenhang zwischen dem Fachwissen der Lehrkräfte und dem Fachwissen ihrer Schülerinnen und Schüler (Blömeke et al., 2010; Döhrmann, Kaiser & Blömeke, 2010).

Allgemein pädagogisches Wissen

Allgemein pädagogisches Wissen meint das Wissen, welches relativ unabhängig von den Fächern für die Optimierung von Lehr-Lernsituationen benötigt wird (Bromme & Rheinberg, 2006). Als Verbindung der Konzepte von Darling-Hammond und Bransford (2005) und Terhart (2002) systematisieren Baumert und Kunter folgende vier Facetten des generischen pädagogischen Wissens und Können: (1) Konzeptuelles bildungswissenschaftliches Grundlagenwissen (Erziehungsphilosophische, bildungstheoretische und historische Grundlagen von Schule und Unterricht, Theorie der Institution, Psychologie der menschlichen Entwicklung, des Lernens und der Motivation), (2) Allgemeindidaktisches Konzeptions- und Planungswissen (Metatheoretische Modelle der Unterrichtsplanung, fachübergreifende Prinzipien der Unterrichtsplanung, Unterrichtsmethoden im weiten Sinne), (3) Unterrichtsführung und Orchestrierung von Lerngelegenheiten (Inszenierungsmuster von Unterricht, effektive Klassenführung, Sicherung einer konstruktiv-unterstützenden Lernumgebung, (4) fachübergreifende Prinzipien des Diagnostizierens, Prüfens und Bewertens (Baumert & Kunter, 2006). Baumert und Kunter strukturieren damit allgemein pädagogisches Wissen anders als Shulman, auf deren Einteilung von Lehrerwissen sie sich eigentlich berufen, und legen das allgemein pädagogische Wissen breiter an. Wissen über Erziehungsziele, bildungstheoretisches und bildungshistorisches Wissen hier zu verorten, folgt dem erziehungswissenschaftlichen Verständnis. Wissen über das Lernen (Gegenstand der pädagogischen Psychologie) nicht als eigene Wissensdomäne aufgefasst deutet hingegen auf die klassische Dreiteilung der universitären Lehramtsausbildung (in zwei Fächer und Erziehungswissenschaften), in der allerdings auch Organisations- und Beratungswissen als erziehungswissenschaftliches Wissen konzipiert sind. Jenes Wissen steht aber nicht in gleichem Maß im Zusammenhang zur Unterrichtstätigkeit wie dies für die anderen Inhalte gilt, die unter Fachwissen, fachdidaktischem Wissen und allgemein pädagogischem Wissen strukturiert wurden.

Dem Projekt TEDS-M liegt eine Vorstellung von allgemein pädagogischem Wissen zugrunde, welche noch stärker auf den Unterricht ausgerichtet ist. Unterschieden werden die vier Anforderungsbereiche (a) Strukturierung von Unterricht (Komponenten- und Prozessbezogene Planung, Analyse von Unterricht, curriculare Strukturierung von Unterricht), (b) Umgang mit Heterogenität (Differenzierungsmaßnahmen, Methodenvielfalt), (c) Klassenführung und Motivation (Störungspräventive Unterrichtsführung, effektive Nutzung der Unterrichtszeit, Leistungsmotivation und Motivierungsstrategien) und (d) Leistungsbeurteilung (Funktionen und Formen, zentrale Kriterien, Urteilsfehler). Dies geschieht im Rückgriff auf eine Vorstellung von „gutem Unterricht“, wie er in Überblicksdarstellungen von Ditton (2000), Good und Brophy (2007) und Helmke (2004) postuliert wird (Blömeke & König, 2010). Auffällig ist eine problemorientierte Struktur im Ansatz von TEDS-M, aber eben auch die damit verbundene Engführung auf das Unterrichtshandeln der Lehrkraft. „Guter Unterricht“ wird als lerneffektiver und leistungsorientierter Unterricht verstanden. Der Bereich der Erziehung sowie außerunterrichtliche Aktivitäten der Lehrkräfte werden (dem Projekt geschuldet) bewusst

ausgeklammert, weil über deren Bedeutung kein internationaler Konsens zu existieren scheint.

Auch Bromme zählt zum pädagogischen Wissen den Teilbereich, bei dem es um Aspekte einer pädagogischen Philosophie geht, folglich den Bereich, der bei Shulman eine eigene Domäne ausmacht. Shulman und Bromme stellen den Erziehungsaspekt deutlicher in den Vordergrund als dies in den auf effektive Wissensvermittlung ausgelegten Konzeptionen von Baumert und Kunter bzw. vor allem Blömeke und König der Fall ist. Darüber hinaus nennt Brumme neben fachlichem, fachspezifischen-pädagogischem (fachdidaktischen) und pädagogischem Wissen curriculares Wissen und die Philosophie des Schulfachs als eigene Wissensdomänen. Während die ersten drei Wissensbereiche klar auf den Unterricht zielen, beinhaltet curriculares Wissen in Brommes Modell auch Aspekte, die vom konkreten Fachunterricht losgelöst sind und sich beispielsweise in Schulprogrammen wieder finden lassen. Er trennt aber im curricularen Wissen und der Philosophie des Schulfachs nicht sauber von individuellen Überzeugungen und Leitlinien der Schule bzw. Fachtraditionen. Die Philosophie des Schulfachs meint bei Bromme auch die Auffassung des Lehrers zu den Fachinhalten und fachspezifischen Methoden sowie epistemologischen Überzeugungen (Bromme, 1997).

Die häufig fehlende Trennung von individuellen Überzeugungen und allgemeinem pädagogischen Wissen wird von Blömeke und König auch grundsätzlich angemerkt. Sie führen dies auf ein enormes Empiriedefizit in diesem Bereich zurück. Die großen Abweichungen in den Konzeptionen können als unterstützender Beleg für die zweite These von Blömeke und König verstanden werden, die das pädagogische Wissen als weniger strukturiert und damit schwieriger zu erfassen charakterisieren (Blömeke & König, 2010). Die Konzeptionen unterscheiden sich darin, ob ein Fokus auf den Unterricht gelegt oder allgemein pädagogisches Wissen sich auf die Lehrertätigkeit im ganzen Kontext Schule bezogen gesehen wird, ob Unterricht auf Lerneffektivität und leistungsorientiert ausgelegt gedacht wird oder ob Erziehungs- und Bildungsziele breiter berücksichtigt werden, und schließlich darin, ob bestimmte Facetten als eigene Wissensdomänen konzipiert und herausgestellt werden. Gemeinsam ist allen hier dargestellten Konzeptionen aber eine äußerst breite Ausdifferenzierung, die die komplexe Aufgabe und die vielen Ansprüche an Lehrkräfte dokumentiert. Jene Ausdifferenzierung ist sicherlich umfangreicher und damit schwieriger fassbar als für das fachwissenschaftliche Wissen und fachdidaktische Wissen.

Beratungswissen und Organisationswissen

Beratungswissen benötigen Experten an den Schnittstellen der Kommunikation mit Laien. Anders als Novizen, die sich am Anfang der Entwicklung zum Experten befinden, sind Laien Personen, die das Wissen und Können der Experten in Anspruch nehmen wollen. Experten müssen nach Rambow und Bromme mit Laien an zwei Stellen kommunizieren: zur Diagnose am Beginn des zu lösenden Problems, um dieses überhaupt erfassen zu können, und nach

dem Bearbeitungsprozess, um dem Laien die Problemlösung zu erklären. Diese Kommunikation wird durch eine unterschiedliche Wahrnehmung des Problems durch Experten und Laien erschwert und dadurch selbst zu einem durch den Experten zu lösenden Problem (Bromme & Rambow, 2001). Experten betrachten Probleme immer in ihrem Expertenkontext. Dies kann dazu führen, dass sie den Laien bei der Diagnose nicht verstehen. Andersherum betrachten Experten auch die Problemlösung aus ihrer Expertensicht, sodass sich die Präsentation der Problemlösung durch den Experten als für den Laien unverständlich darstellen kann. Dieses für den Kommunikationsprozess benötigte Wissen entwickelt sich einerseits unabhängig von domänenspezifischem Wissen und problembezogenem Können im eigentlichen Sinne, ist aber doch gleichzeitig eng mit diesem verbunden (Rambow & Bromme, 2000). Bromme und Rambow bemängeln zur Expertiseforschung, dass in dieser die Experten-Laien-Kommunikation meistens allerdings ausgeblendet werde (Bromme & Rambow, 2001).

Auf die Schulsituation übertragen kann dies bedeuten, dass Schüler und Schülerinnen und Erziehungsberechtigte als Laien aus Sicht der Profession für das Lehren und Lernen verstanden werden. Dies schließt umfassende Lern- und Entwicklungsprozesse ein, welche in Eltern- oder Schülergesprächen thematisiert werden (Hertel & Bruder, 2009). Hertel u.a. bilanzieren eine geringe Gesprächsbereitschaft von Lehrkräften mit Erziehungsberechtigten als tatsächliche Beratung (stattdessen wird Mitteilung bevorzugt) und führen dies auf mangelnde Ausbildung der Lehrkräfte zurück (Hertel & Bruder, 2009). Hertel und Kollegen setzen den Rahmen für Beratung allerdings enger als von der Begriffserklärung durch Rambow und Bromme nötig: Für Rambow und Bromme ist das Erklären der Medikation durch den Arzt beratende Kommunikation (Rambow & Bromme, 2000). Dementsprechend kann schon das Erklären der Unterrichtsziele als Beratung der Schülerinnen und Schüler durch die Lehrkraft angesehen werden. Eine Lehrkraft befindet sich aber auch in der Beraterrolle, wenn sie mit Erziehungsberechtigten bzw. Schülerinnen und Schülern die Ergebnisse von Schulleistungsmessungen bespricht, denn auch für deren Deutung ist sie als Expertin anzusehen. Dazu benötigt die Lehrkraft allerdings fachdidaktisches Wissen, spezieller diagnostisches Wissen, und allgemein pädagogisches, spezieller statistisches Methodenwissen. Beratungswissen im weiteren Sinne einer Beratungskompetenz ist folglich auch eng mit den anderen Wissensbereichen verwoben. Lehmann-Grube und Nickolaus schlagen daher vor (mit Blick auf den eingeschränkten Anwendungsbereich bei Hertel und Kollegen nachvollziehbar), Beratungswissen als Teil des erziehungswissenschaftlichen Wissens zu verorten (Lehmann-Grube & Nickolaus, 2009).

Unter Organisationswissen wird Wissen über den schulischen Kontext verstanden, beispielsweise Wissen über Bildungspolitik, Bildungsadministration, Bildungsfinanzierung, Wissen über die Organisationsstruktur von Schule, Managementwissen (Shulman, 1986, 1987). Lehrkräfte müssen aber nach Fried auch über Wissen verfügen, welches Organisations-, Unterrichts- und Personalentwicklung ermöglicht. Darunter fasst sie Wissen über die Nutzung von Selbsterfahrung, Supervision, kollegiale Beratung und Fallstudien (Fried, 2002). Letzteres könnte (und soll im Folgenden) auch als Teil eines

Schulentwicklungswissens begriffen werden. Dieses muss (wie unten ausgeführt wird) auch Teile der anderen vier vorher skizzierten Wissensdomänen beinhalten, liegt also quer zu den fünf Dimensionen, die von Baumert und Kunter genannt werden.

4.2.2 Kausal- und Zielüberzeugungen als gegenstandsbezogene Überzeugungen

Eine theoretische Differenzierung

Der zweite wichtige Bestandteil eines Lehrer-Handlungskompetenzmodells sind Kausal- und Zielüberzeugungen. Wissen und Können allein wird nicht als handlungswirksam angesehen, sondern bedarf der Ergänzung durch Einstellungen, Wertvorstellungen und Ziele. Es muss aber vor allem in handlungswirksame Kognitionen umgewandelt werden. Wie oben bereits dargestellt, ist das Denken der Lehrkraft dem Lehrer-Expertenmodell folgend ein zentraler Aspekt (Bromme, 1997). Verschiedene Studien haben sich mit der Bedeutung so genannter Überzeugungen für das Lehrer-Handeln beschäftigt. Nach Brunner und Kollegen zeigt die bisherige Forschung zu Überzeugungen von Lehrkräften (Teacher Beliefs), dass Überzeugungen über die Struktur des Wissensgebiets, über das Lehren und Lernen und über Aufgaben der Profession sowohl langfristig als auch im unmittelbaren Unterrichtskontext handlungssteuernde Funktion haben können und Lehrkräfte diese im Verlauf ihres beruflichen Lebens dezidiert erwerben (Brunner, Kunter, Krauss & Klusmann et al., 2006). Allerdings ist die empirische Befundlage über den Zusammenhang von Überzeugungen und Unterrichtsqualität eher inkonsistent (Leuchter, Reusser, Pauli & Klieme, 2008).

Ein Grund hierfür mag der Begriff “Überzeugung“ selbst sein. Parallel zu den Wissensdomänen zeigt sich auch hier ein Begriffswirrwahr. Dem Begriff “Überzeugung“ wird sich auf verschiedene Weise genähert und die zugrunde liegenden Vorstellungen über den Begriff sind wenig kongruent. Damit hängt zusammen, dass bisher nicht geklärt ist, mit welchem Verfahren Überzeugungen zu messen sind (Leuchter et al., 2008). Stellvertretend sollen drei Anordnungsmöglichkeiten hier dargestellt werden: den inhaltlich-funktionalen Zugang von Baumert und Kollegen, den anforderungsorientierten Zugang nach Blömeke und Kollegen und den wirkungsmächtigen Zugang nach Leuchter und Kollegen. Letzterem ähnlich ist der Zugang nach Dann, der für den Begriff der “subjektiven Theorie“ Klärung bringt. Alle hier vorgestellten Zugänge können das Verhalten von Lehrkräften erklären. Die jeweils berichteten Ergebnisse sind aber erst vergleichbar, wenn die Zugänge in Beziehung gesetzt werden.

In Anlehnung an Fenstermacher (1994) sprechen Baumert und Kunter bei Überzeugungen von (1) Wertbindungen, konkret der Berufsmoral der Lehrkräfte, in deren Zentrum die Verpflichtung auf Fürsorge im Sinne von Fördern und Fordern, Gerechtigkeit in seinen verschiedenen Facetten und Wahrhaftigkeit steht, von (2) epistemologischen

Überzeugungen über die Struktur und Genese des Fachwissens, von (3) subjektive Theorien über das Lehren und Lernen und von (4) Zielsysteme für Curriculum und Unterricht.

Diese vier Kategorien vermischen inhaltliche und funktionale Unterschiede. Sowohl Wertbindungen als Zielsysteme für Curriculum und Unterricht können als Ziele aufgefasst werden. Sie sind funktional ähnlich und könnten als normative Überzeugungen bezeichnet werden. Während die Berufsmoral der Lehrkraft aber allgemeinere Ziele für die komplette Lehrertätigkeit definiert, sind die Zielsysteme für Curriculum und Unterricht spezifischer auf das Kerngeschäft der Lehrkräfte, den Unterricht, ausgerichtet. Diese Abstufung findet sich auch bei den beiden anderen Kategorien. Epistemologische Überzeugungen und subjektive Theorien über das Lehren und Lernen gehören in die Gruppe der Überzeugungen, die das Funktionieren der Welt erklären. Überzeugungen beider Kategorien sind als Wenn-dann-Sätze konstruierbar, weil epistemologische Überzeugungen auch immer Vorstellungen darüber sind, wie Wissen gelernt und gelehrt werden kann („Wer Mathematik verstehen will, muss ausdauernd puzzeln."). Beide Kategorien beinhalten aber gleichzeitig auch vorkausale Überzeugungen, nämlich dann, wenn es um Überzeugungen geht, die noch keine Handlungsweise nahe legen („Üben hilft immer.").

Blömeke und Kollegen wählen einen anforderungsorientierten Zugang. Dieser besitzt ebenfalls vier Kategorien, diese sind aber nicht funktional verschieden, sondern besitzen sowohl normative als auch erklärende Elemente. Sie ersetzen die Wertvorstellungen durch professionsbezogene Überzeugungen und sie erweitern die Gruppe der Überzeugungen darüber hinaus um selbstbezogene Überzeugungen. Darunter verstehen sie (beispielsweise) Selbstwirksamkeitswahrnehmungen (SWE) und Berufsmotivation. Selbstbezogene Überzeugungen haben im COACTIV-Modell ein Äquivalent in der motivationalen Orientierung und der Selbstregulationsfähigkeit (s.u.). Insgesamt werden bei MT21, der Pilotstudie zu TEDS-M [78], von Blömeke, Kaiser und Lehmann folglich diese vier Überzeugungen unterschieden: epistemologische Überzeugungen, unterrichtsbezogene Überzeugungen, professionstheoretische Überzeugungen und selbstbezogene Überzeugungen (Blömeke, Müller, Felbrich & Kaiser, 2008). Es handelt sich bei diesen Kategorien nicht nur um vier Anforderungsbereiche, es sind auch vier verschiedene konkrete Ebenen herausgearbeitet. Dadurch werden auch Überzeugungen angesprochen, die den verengten Bereich des Unterrichtens verlassen.

Das dritte Lehrer-Handlungskompetenzmodell stammt aus dem Projekt „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis" (Klieme & Reusser, 2003). Auch Leuchter, Pauli, Reusser und Klieme gehen bei Lehrkräften von einer Wissensbasis aus. Diese Wissensbasis konstruieren sie als eine Kombination aus systematisch im akademischen Umfeld erworbenem deklarativen Wissen (Wissen der Unterrichtsfächer, pädagogisch-psychologisches und fachdidaktisches Wissen) und einem in der Praxis geformten

[78] Für TEDS-M wurde nur die Dimension der epistemologischen Überzeugungen übernommen und dort wurde zwischen Überzeugungen über Struktur von Mathematik und Überzeugungen über den Erwerb von mathematischem Wissen unterschieden.

Erfahrungs- und Reflexionswissen (Leuchter, Pauli, Reusser & Klieme, 2006). Leuchter und Kollegen nehmen grundsätzlich an, dass Überzeugungen das Lehrerhandeln beeinflussen können, unterscheiden aber nach Alisch zwischen verhaltensfernen Kognitionen, zu denen sie subjektive Theorien zählen und die durch Fragebögen oder Strukturlegetechniken wissenschaftlich erhoben werden, und verhaltensnahen Kognitionen, die retrospektiv beispielsweise über lautes Denken einer Lehrkraft bei der nachträglichen Ansicht einer ihrer Unterrichtsstunde gemessen werden. Im Gegensatz zu verhaltensnahen Kognitionen sind verhaltensferne Kognitionen nicht zwingend in konkreten Situationen handlungsleitend. Ihnen kann aber zugestanden werden, auf der Planungsebene zu wirken und somit indirekt handlungsleitend zu sein, wenn es nicht zu unvorhersehbaren Ereignissen kommt. Verhaltensferne Kognitionen sind folglich globaler (Leuchter et al., 2008). Weiter wird davon ausgegangen, dass der Grad der SWE und die beruflich erlebte Belastung darüber entscheiden, wie schnell verhaltensferne und verhaltensnahe Kognitionen in einer konkreten Situation divergieren (Leuchter et al., 2006). Das implizit zugrunde gelegte Modell besitzt also ebenfalls eine motivationale und selbstbezogene Komponente.

Dann listet persönliche Konstrukte, semantische Netzwerke, praktisches Wissen und Fallwissen neben subjektiven Theorien von Lehrkräften und sieht diese im Kontext konstruktivistischer Theorien als Wissensrepräsentationen. Wenngleich er erziehungs- und schulbezogene Einstellungen und Werthaltungen von Lehrkräften als Gegenstand der sozialpsychologischen Forschung ebenfalls in diese Liste aufnimmt (Dann, 2008), fokussiert seine Auflistung auf das Verhältnis von Wissensrepräsentation aus wissenschaftlicher Forschung und beruflicher Erfahrung des Praktikers auf der einen Seite und anderen Einstellungen wie Wertvorstellungen auf der anderen Seite. Wissensrepräsentationen und Einstellungen anderer Art sollen im Folgenden unterschieden werden. Zentral ist hierbei der Begriff der “subjektiven Theorie“.

In Anlehnung an den Begriff der wissenschaftlichen (als von der Gesellschaft geteilt angenommenen) Theorie sollen Wissensrepräsentationen als “subjektive Theorien“ bezeichnet werden: „Subjektive Theorien sind komplexe Formen der individuellen Wissensorganisation. […] Subjektive Theorien enthalten […] Wissenselemente (inhaltliche Konzepte), die in bestimmten Beziehungen (formale Relationen) zueinander stehen, so dass Schlussfolgerungen möglich sind […]. [Sie] erfüllen analog [zu den wissenschaftlichen Theorien] für den Alltagsmenschen „…die Funktionen (a) der Situationsdefinition i.S. einer Realitäts-Konstituierung, (b) der nachträglichen Erklärung […] eingetretener Ereignisse, (c) der Vorhersage […] künftiger Ereignisse, (d) der Generierung von Handlungsentwürfen oder Handlungsempfehlungen zur Herbeiführung erwünschter oder zur Vermeidung unerwünschter Ereignisse“ (Dann 1994: 166). Darüber hinaus kommt zumindest bestimmten subjektiven Theorien eine handlungsleitende und handlungssteuernde Funktion zu“ (Dann, 2008).

Bei subjektiven Theorien können mit Herstellungswissen (prozedurales Wissen zur Auswahl aus Handlungsalternativen) und Funktionswissen (deklaratives Wissen über das

Zustandekommen von psychischen Ereignissen) mindestens zwei Wissensarten unterschieden werden. Funktionswissen kann durch häufiges Handeln zu Herstellungswissen verdichtet werden, beide können sich aber auch unabhängig voneinander weiter entwickeln. Subjektive Theorien werden von Lehrkräften zur Handlungssteuerung herangezogen und können als Wissensbasis bezeichnet werden (Dann, 2008). Diese Unterscheidung umschließt somit die beiden Varianten der oben als „erklärende Überzeugungen" bezeichneten Überzeugungs-Kategorien nach dem COACTIV-Modell und differenziert hier begrifflich klarer. Darüber hinaus wird bei Dann noch einmal die Überlappung von Wissen und Können als getrenntes Konstrukt von Überzeugungen deutlich. Weil Wissen und Können in das subjektive semantische Netzwerk integriert werden muss, um handlungswirksam sein zu können, scheint diese Unterscheidung unter konstruktivistischen Annahmen nicht disjunkt, sondern viel mehr künstlich. Auch Sembill und Seifried sprechen davon, dass objektives Wissen und subjektive Sichtweisen nicht kategorial zu trennen sind. Eine Trennung von kognitiven, emotionalen und motivationalen Elemente scheint ihnen nicht Ziel führend, da jede Verarbeitung von „Fakten" unter bestimmten emotionalen Bedingungen im sozialen Kontext geschieht (Sembill & Seifried, 2009). Diesbzgl. unterstreichen Baumert und Kunter den Unterschied zwischen Wissen und Überzeugungen: Überzeugungen sind stets subjektiv und entziehen sich dadurch einem wissenschaftlichen Diskurs. Sie werden (im strengen Sinne) nicht gerechtfertigt und validiert, sodass Menschen über subjektive Theorien verfügen können, die logisch kollidieren (Baumert & Kunter, 2006). Demgegenüber sind subjektiv verarbeitete wissenschaftliche Theorien Wissen auf Basis wissenschaftlicher Forschung.

Resümieren wir über die vier vorgestellten Zugänge, lässt sich folgendes Bild festhalten: (1) Unterschieden werden muss zwischen verhaltensfernen Kognitionen und zwischen verhaltensnahen Kognitionen. Verhaltensferne Kognitionen sind überdauernd und allgemeinerer Natur als verhaltensnahe Kognitionen. Verhaltensnahe Kognitionen sind stets die handlungsleitenden Kognitionen, müssen aber nicht immer mit den verhaltensfernen Überzeugungen kongruent sein. (2) Verhaltensferne Kognitionen lassen sich dadurch mittels Fragebogen erfassen. (3) Bei den verhaltensfernen Kognitionen finden wir einerseits Überzeigungen als Repräsentationen von Wissen (subjektive Theorien), welche Kausalstrukturen gehorchen, ohne dabei untereinander zwingend widerspruchsfrei zu sein, und andererseits Überzeugungen in Form von Einstellungen, Wertvorstellungen und Zielen.

Nachfolgend seien an dieser Stelle einige Befunde zu Überzeugungen als handlungswirksame Kognitionen für das Lehrerhandeln im Unterricht vorgestellt, um deren Bedeutung, aber auch die in diesem Forschungsgebiet gegebenen Schwierigkeiten aufzuzeigen. Anschließend werden zwei Arten von Zielüberzeugungen vorgestellt, die für Unterrichtshandlungen von Lehrkräften relevant sind.

Handlungsleitende Kognitionen

Leuchter und Kollegen haben im Projekt „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis“ (Klieme & Reusser, 2003) u.a. die Zusammenhänge von (subjektiven) Überzeugungssystemen einerseits und handlungsleitenden Kognitionen andererseits sowie dem Unterrichtshandeln untersucht. In ihrer Untersuchung konnten sie die ursprünglich angenommene Funktion von verhaltensfernen Kognitionen aber nur für deutsche Lehrkräfte zeigen, die ein rezeptives Unterrichtsverständnis haben. Bei einem konstruktivistischen Unterrichtsverständnis korrelierten (bei Kontrolle der SWE und belastenden Faktoren) verhaltensferne und verhaltensnahe Kognitionen weit weniger stark (Leuchter et al., 2006). Für diesen scheinbar widersprechenden Befund lassen sich verschiedene Erklärungen finden. Leuchter und Kollegen bieten mit Verweis auf Bromme an, dass die meisten Handlungen Routinehandlungen sein könnten und daher in den Interviews, mit denen die verhaltensnahen Kognitionen erfasst werden sollten, nicht genannt wurden (Leuchter et al., 2006). Dies widerspricht allerdings dem Befund von Diedrich und Kollegen. Verhaltensferne Kognitionen können je nach Situation in verhaltensnahe übergehen. Als andere Begründung kann auf die kleine Stichprobe von nur 20 Lehrkräften insgesamt verwiesen werden (Leuchter et al., 2008). Evtl. ist es aber in der Tat so, dass konstruktivistische Überzeugungen in abstrakter, verhaltensferner Form schwieriger zu konkreten verhaltensnahen Kognitionen transferierbar sind und diese Überzeugungen nicht entsprechend verinnerlicht sind. Es ist Leuchter und Kollegen zuzustimmen, wenn sie für die Wirkung von Überzeugungen auf das Verhalten noch viele offene Fragen und Forschungsbedarf sehen.

Epistemologischen Überzeugungen: Transmitterorientierung und Konstruktionsorientierung

Dubberke und Kollegen konnten einen Zusammenhang zwischen so genannten Transmitterüberzeugungen (Stichwort „Nürnberger Trichter“) und der Unterrichtsqualität (in diesem Fall: repetitives Üben bevorzugt, geringeres kognitiv herausforderndes Unterrichtsniveau) finden. Dies sind epistemologische Überzeugungen, die mit subjektive Theorien über das Lehren und Lernen und Unterrichtsziele zu Überzeugungssyndromen kombiniert werden (Dubberke et al., 2008). Auch Schmotz u.a. berichten, dass Lehrkräfte mit Transmitterorientierung einen gerichteten Vermittlungsprozess und ausgedehnte Übungsphasen, Lehrkräfte mit Konstruktionsorientierung hingegen einen schülerorientierten Unterricht und selbstgesteuerten Lernprozess bevorzugen (Schmotz, Felbrich & Kaiser, 2010). Leuchter und Kollegen konnten den umgekehrten Befund, eine konstruktivistische Sicht auf den Mathematikunterricht, als maßgeblichen Faktor für herausfordernden, verknüpfenden Unterricht, nicht replizieren (Leuchter et al., 2008). Schmotz, Felbrich und Kaiser berichten von einer höheren Zustimmung zur Konstruktionsorientierung und niedrigerer Zustimmung zur Transmitterorientierung von angehenden Gymnasiallehrkräften im Vergleich zu anderen angehenden Mathematiklehrkräften in TEDS-M (Schmotz et al.,

2010). Einen Zusammenhang zwischen unterrichtsbezogenen Kognitionen und (selbstberichtetem) Instruktionsverhalten fanden hingegen auch Diedrich und Kollegen. Auch scheinen Überzeugungen über die Gestalt des Unterrichtsgegenstands (Mathematik als Prozess) und den Unterrichtszielen (hohe Schüleraktivierung) vermehrt gemeinsam aufzutreten. Lehrkräfte sahen sich außerdem eher für den Unterrichtserfolg verantwortlich, wenn sie zugleich einen prozessorientierten, Schülerinnen und Schüler aktivierenden Unterricht bevorzugten (Diedrich et al., 2002).

Unterrichtsziele

Allgemein können Ziele als Repräsentation eines gewünschten Verhaltens definiert werden. Durch Zielsetzungen werden Diskrepanzen zwischen einem Ist- und einem Sollzustand erzeugt. Zielsetzungen können als Verbindlichkeit gegenüber dem Erreichen des erwünschten Ereignisses oder Verhaltens aufgefasst werden. Ziele können von einer Person selbst gesetzt werden, dann kann dies über positive Erwartungen an die Zukunft und Vermeidung von gegenwärtigen Hindernissen zu großem Engagement führen. Daneben existieren aber auch aufgetragene Ziele. Der Grad, mit dem eine Person diese als für sie verbindlich ansieht und der ihr Engagement für das Erreichen der Ziele bestimmt, hängt von den dadurch erwarteten Schwierigkeiten des Ziels, der Beteiligung beim Setzen des Ziels, der Vertrauenswürdigkeit der zielsetzenden Person oder des Instituts und extrinsischer Belohnung ab (Sevincer & Oettingen, 2009).

Bromme und Rheinberg (2006) nehmen an, dass der Erfolg von Lehrkräften durch die Ziele beeinflusst wird, die sie erreichen wollen. Studien demonstrieren aber häufig eher eine Diskrepanz zwischen geäußerten Zielen und tatsächlichem Verhalten. Bromme und Rheinberg erklären dies durch die Aufgabenfülle, die von Lehrkräften gleichzeitig erfüllt werden müssen. Allgemeinere Ziele treten bei konkurrierenden kurzfristigen Zielen, die durch störendes Schülerverhalten oder andere Störungen hervorgerufen werden, oft in den Hintergrund. Allgemeine Ziele werden nach Köttl und Sauer vorwiegend dann handlungswirksam, wenn sich Lehrkräfte in „lernbereiten Klassen“ befinden (1980 nach Bromme & Rheinberg, 2006). Nach der Zielsetzungstheorie spielen verschiedene Moderatorvariablen bei der Umsetzung von Zielen eine Rolle: Neben der benötigten aufgabenspezifischen Fähigkeit sind dies die Selbstwirksamkeitserwartung, die schon angesprochene Zielbindung, gegebene Rückmeldungen und die Aufgabenkomplexität (Beispiel „schwierige Klassen“) (Spörl, 2009; Wegge & Schmidt, 2009).

Bei den Zielen für den Unterricht lassen sich theoretisch und empirisch kognitive und affektive-motivationale (Lern-)Ziele unterscheiden (Müller, Felbrich & Blömeke, 2008). Lernziele sind erst einmal in doppelter Weise fremdgesetzte Ziele, die durch die Bildungspolitik vorgegeben und beispielsweise mittels Lehrplänen kommuniziert werden. Sie werden ggf. allerdings zu selbstgesetzten Lehr-Zielen, wenn Lehrkräfte sie im Rahmen ihrer Autonomie in ihren Zielkatalog integrieren. Zu tatsächlichen Lernzielen können sie erst

werden, wenn Schüler und Schülerinnen diese für sich adaptieren bzw. das Verhalten zeigen, welches zum Erreichen der Lernziele notwendig ist.

Für kognitive Ziele, die man als inhaltliche Lernziele (eigentlich besser Lehrziele) bezeichnen könnte, existiert für Mathematik ein grober Konsens über ihren Umfang. Man findet die Unterrichtsziele beispielsweise in den Bildungsstandards für Mathematik der KMK (2004 - dort als Leitideen und allgemeine Kompetenzen bezeichnet) und in den Standards des National Council of Teachers of Mathematics (2000). Über die Bildungsstandards haben sie auch Eingang in die Kernlehrpläne Nordrhein-Westfalens gefunden, sind dort aber anders gegliedert (vier Inhaltskompetenzen statt fünf Leitideen und vier Prozesskompetenzen statt sechs allgemeiner Kompetenzen). In MT21 wurden angehende Mathematiklehrkräfte nach der Relevanz von Lernzielen befragt. Es konnte eine Vier-Faktoren-Struktur für erstrebenswerte Lernziele im Sinne der allgemeinen Kompetenzen bzw. Prozesskompetenzen nachgewiesen werden. Als klar abgrenzbar von anderen Faktoren zeigte sich der Faktor „Routineaufbau“, „Problemlösen und Modellieren“ und „Beweisen“ wiesen ebenfalls nur eine Korrelation von r.13 auf. „Argumentieren und Begründen“ hingegen korrelierte mit jenen beiden Faktoren aber im mittleren Bereich (r.38 bzw. r.46). Dies könnte darauf deuten, dass diese Kategorie keine unabhängige Kategorie ist. Konsequenterweise sind Beweisen, Argumentieren und Begründen in den Kernlehrplänen Nordrhein-Westfalens auch eine gemeinsame Kompetenz. Hingegen stellen Problemlösen und Modellieren dort jeweils eigene Kompetenzen dar. Die Skala wies mit .55 bei MT21 auch einen eher wenig befriedigenden Wert auf (welches aber vorwiegend auf die Größe der Skala zurückgeführt wird Müller et al., 2008).

Den affektiven-motivationalen Lernzielen liegt die Vorstellung eines Unterrichts als inszenierte Lerngelegenheiten zugrunde (Rheinberg & Salisch, 2008; Schiefele & Krapp, 2000). Eine hohe Motivation führt zu einem stabilen hohen Interesse an einem Inhaltsgebiet und dadurch indirekt zu höheren Schülerleistungen (Krapp 2001 nach Müller et al., 2008). Interesse ist darüber hinaus immer gefühlsbezogen und mit Wertvorstellungen verknüpft (Deci, Ryan, 1985 nach Müller et al., 2008).

Sowohl bei den affektiven-motivatiobalen als auch bei den kognitiven Lernzielen fanden Müller u.a. eine sehr hohe Zustimmung bei Referendaren. Die Lernziele „Beweisen“ erhielt dabei über alle beteiligten Schulformen die geringste Zustimmung der kognitiven Lernziele. „Problemlösen und Modellieren“ erzielte den größten Zustimmungswert (Müller et al., 2008). Letzteres überrascht, da die ähnlich angelegte Dimension „Mathematisierungs- und Modellierungsfähigkeit“ in COACTIV eine geringere Zustimmung erfahren hatte (Baumert et al., 2004). Müller und Kollegen deuten dies als einen möglichen Effekt der veränderten Lehramtsausbildung (Müller et al., 2008).

4.2.3 Berufserleben als personenbezogene Überzeugungen

Eine Einteilung in Überzeugungen und Fähigkeiten

Anders als die bisher unter dem Namen „gegenstandsbezogene Überzeugungen“ erläuterten Überzeugungen, die sich über Zusammenhänge in Bezug auf Unterrichtsinhalte und Unterrichtziele beziehe, sind „personenbezogene Überzeugungen“ kognitive Strukturen, die Überzeugungen über die eigene Person darstellen, in ähnlicher Form auch in anderen Berufsfeldern existieren und bis zu einem gewissen Grad beispielsweise als vom Unterrichtsfach unabhängig angesehen werden können. Im COACTIV-Modell sind diese in motivationale Orientierung und Selbstregulation eingeteilt. Motivationale Orientierung und Selbstregulation spielen nach Baumert und Kunter eine wesentliche Rolle für die psychische Funktionsfähigkeit von handelnden Personen. Sie regeln bzw. überwachen das berufliche Handeln und halten die Intention aufrecht. Forschungstypologisch finden sich vor allem zwei Stränge, nämlich vornehmlich Arbeiten zur Selbstwirksamkeit sowie Arbeiten zum Belastungserleben und schützenden Faktoren im Lehrerberuf (Baumert & Kunter, 2006). Berliner zählt die Regulierungsfähigkeit zu einer der Variablen, in denen sich Novizen und Experten unterscheiden lassen (Berliner, 2001). Allgemein bedeutet Selbstregulation die Kenntnis eines Soll-Zustands, die Messung eines Ist-Zustands und die Möglichkeit, eine Differenz in Richtung des Soll-Zustands durch gezielte Maßnahmen zu reduzieren (Schmitz & Schmidt, 2007). Darüber hinaus werden aber bei menschlichen (oder besser: nicht automatischen) Selbst-Regulations-Prozessen Fähigkeiten und die Bereitschaft zur Messung des Ist-Zustands, die Anerkennung des Soll-Zustands und die Bereitschaft zu Maßnahmen in Richtung des Soll-Zustands benötigt. Selbstregulation kann in verschiedenen Kontexten verstanden werden, u.a. Selbstregulation des Lernens (vgl. (Landmann & Schmitz, 2007), s.a. (4.3)). Selbstregulation kann aber eben auch wie bei Baumert und Kunter im Sinne eines Ressourcenmanagements verstanden werden. Der entsprechende Verarbeitungsprozess beruflicher Eindrücke und der beruflichen Situation wird dann als Coping bezeichnet (Dick, 2006).

Die Einordnung von „Kontrollüberzeugungen und Selbstwirksamkeitserwartungen“ als „psychologische Funktionsfähigkeit“ (Baumert & Kunter, 2006) suggeriert Handlungsoptionen oder verfügbare Werkzeuge im Sinne anderer Fähigkeiten (beispielsweise die mathematische Fähigkeit des Problemlösens), die von den üblicherweise zur Messung eingesetzten Instrumenten nicht gegeben sind. Erfasst wird, ob die Personen glauben, diese Fähigkeit zu besitzen. Auch handelt es sich dabei nicht direkt um eine Form der Motivation, wenngleich es diese als sekundären Effekt aufrechterhalten kann. Daher erscheint eine Einordnung wie von Blömeke, Kaiser und Lehmann als eine Kategorie der Überzeugungen (selbstbezogene Überzeugungen) nachvollziehbar.

Anders verhält es sich bei den ressourcenbezogenen Kognitionen wie sie in den Konzepten zum Arbeitsengagement, der Widerstandsfähigkeit und dem beruflich erlebten

Belastungserleben konkretisiert sind. Hierin spiegeln sich motivationale Elemente (besonders deutlich im Arbeitsengagement), Fähigkeiten (besonders deutlich in der Widerstandsfähigkeit) und affektive Elemente (erlebte berufliche Belastung) wider, die stark miteinander verschlungen sind (s.u.). Gleichzeitig werden hier Überzeugungen erfasst, die sowohl eigenschaftlich sind („Bei der Arbeit kenne ich keine Schonung." i.S. von „Ich bin stark motiviert bei der Arbeit.") als auch auf Fähigkeiten bezogen sind („Ich neige dazu, über meine Kräfte hinaus zu arbeiten." i.S. von „Ich kann meine Kräfte nicht einteilen.") (Beispielitems aus COACTIV zum Arbeitsengagement Baumert et al., 2008). Sowohl die Selbstwirksamkeitserwartung als auch die im Bereich des Ressourcenmanagements zusammengefassten Konzeptionen sind durchaus als Überzeugungen aufzufassen. In Abgrenzung zu subjektiven Theorien, Zielen und Werthaltungen sollen die hier angesprochenen Überzeugungen als personenbezogene Überzeugungen bezeichnet werden. Zuerst wird auf persönliche Überzeugungen im Kontext des beruflichen Beanspruchungserlebens eingegangen. Unter dem Konstrukt persönliche Überzeugungen im Kontext des beruflichen Beanspruchungserlebens werden die drei Konstrukte Arbeitszufriedenheit/Arbeitsengagement, Belastungserleben und Overcommitment vorgestellt und mit Blick auf die Lehrergesundheit und die Folgen für den Unterricht diskutiert. Anschließend werden Aspekte zum Fähigkeitsselbstkonzept von Mathematiklehrkräften und der Selbstwirksamkeit erläutert.

personenbezogene Überzeugungen im Kontext des beruflichen Beanspruchungserlebens

Die gesundheits- und arbeitswissenschaftliche Forschung zur Beziehung von Gesundheit und beruflicher Leistung offenbart einen wechselseitigen Einfluss. Gesundheit und Wohlbefinden, bemerken Schumacher, Paulus und Sieland (2009), sind wichtige Voraussetzungen, um Leistungspotenziale zu realisieren. Erlebte Wirksamkeit des eigenen Handelns und beruflicher Erfolg stärken aber auch umgekehrt das Wohlbefinden und infolgedessen die Gesundheit (Schumacher, Paulus & Sieland, 2009). Baumert und Kunter bilanzieren nach Sichtung von Studien von Maslach, Schaufeli, Leiter (2001) sowie Rudow (1999) und genauso Schaarschmidt (2002) sowie Klusmann (2006), dass insbesondere der verantwortungsvolle Umgang mit den eigenen persönlichen Ressourcen ein wichtiger Teil der allgemeinen pädagogischen Handlungskompetenz ist (Baumert & Kunter, 2006). Auch Schaarschmidt spricht davon, dass die Anforderungen nicht nur unter dem Aspekt der Vermittlung von Lehr- und Erziehungsprozessen zu betrachten sind, sondern der Blick auch auf die Auswirkungen der Lehrerarbeit auf die Betreffenden selbst zu richten ist. Ohne Lehrergesundheit ist eine hohe Qualität des Lehrens und Lernens auf Dauer nicht denkbar (Schaarschmidt, 2009). Lehrkräfte müssen folglich nicht nur Experten für das Lehren und Lernen, sondern auch im Umgang mit den eigenen Ressourcen sein.

Identifiziert werden können hierbei als Variablen Arbeitsengagement (Bedeutsamkeit der Arbeit, beruflicher Ergeiz, Verausgabungsbereitschaft, Perfektionsstreben), Widerstandsfähigkeit (Distanzierungsfähigkeit, Resignationstendenz bei Misserfolg, offensive

Problembewältigung, innere Ruhe und Ausgeglichenheit), beruflich verbundene Emotionen (Arbeitszufriedenheit, Commitment, Erfolgserleben im Beruf) und Beanspruchung (Böhm-Kasper, 2004; Böhm-Kasper, Bos, Körner & Weishaupt, 2001; Dick, 2006; Schaarschmidt, Kieschke & Fischer, 1999; Schaufeli & Bakker, 2003). Die Forschung bietet dazu insbesondere aus den letzten fünfzehn Jahren einige interessante Befunde zu Lehrkräften. Allerdings kann bzgl. des Lehrberufs eindeutig von einer negativen arbeits- und organisationspsychologischen Forschung gesprochen werden: Fast alle Studien beschäftigen sich mit der Problematik der nicht gesunden Lehrkräfte, während nur ca. 5% der Studien die Arbeitszufriedenheit und Arbeitsmotivation thematisieren (Klusmann, 2008a; Schaufeli & Bakker, 2003).

Die Konzeption dieser Arbeit liegt auf den drei Bereichen Arbeitszufriedenheit/Arbeitsengagement, Belastung/Beanspruchung und Commitment, welche auch von Klusmann als zentrale Bausteine der arbeits- und organisationspsychologischen Forschung zum Beruf Lehrer eingestuft werden (Klusmann, 2008a).

Arbeitszufriedenheit und Arbeitsengagement

Arbeitszufriedenheit verbindet die emotionale Reaktion auf die Arbeit (oder die Meinung über die Arbeit) mit der Bereitschaft, sich in der Arbeit in bestimmter Weise zu verhalten (Nerdinger, Blickle & Schaper, 2008). In der arbeits- und organisationspsychologischen Forschung kann die Arbeitszufriedenheit unter zwei Gesichtspunkten verstanden werden. Einmal ist Arbeitszufriedenheit eine Moderatorvariable für die Arbeitsleistung bzw. das Arbeitsengagement. Der Zufriedenheit des Arbeitenden kommt die Rolle eines Mittels zur Leistungssteigerung zu. Aus Forschungssicht stellen die Zufriedenheit der Menschheit und damit die Zufriedenheit des Arbeitenden mit seiner Tätigkeit aber auch einen Wert an sich dar.

Obwohl an Lehrkräfte sehr große Anforderungen gestellt werden und diesbzgl. eher eine geringe Arbeitszufriedenheit zu erwarten wäre, lässt sich diese Erwartungen in Studien nicht derart eindeutig bestätigen. Die Tendenz ist sogar eher positiv (Dick, 2006). Bromme und Rheinberg führen dies auf die Arbeitsinhalte (Gestaltungsmöglichkeiten, erlebte Verantwortung für die Entwicklung von Menschen) des Lehrerberufs zurück. Die Bezahlung oder Arbeitszeiten spielten weniger eine Rolle (Nerdinger et al., 2008). Auch fehlt es häufig an direkten, bestärkenden Rückmeldungen durch Kollegen, Vorgesetzte oder Schülerinnen und Schüler bzw. Erziehungsberechtigten (Dick, 2006). Zwar bewerten Lehrkräfte positive Rückmeldungen als die wichtigste Art der Belohnung für ihre Arbeit (noch vor der Bezahlung und der Jobsicherheit bzw. Aufstiegsmöglichkeiten) (Lehr, Hillert & Keller, 2009), doch scheint dies insgesamt nur eine untergeordnete Rolle zu spielen. Gerade die große Herausforderung steigere bei vielen Lehrkräften die Zufriedenheit mit ihrer Tätigkeit (Bromme & Rheinberg, 2006). Wichtig ist das richtige Verhältnis der Aufgabenschwierigkeit

zu den selbst wahrgenommenen Fähigkeiten. Dementsprechend unterscheiden sich hier auch Experten und Novizen. Experten fühlen sich unabhängiger als Novizen und sind aufgrund größerer Handlungsoptionen zufriedener mit ihrer Expertentätigkeit (Berliner, 2001).

Eine hohe Arbeitszufriedenheit sorgt aber auch für mehr Motivation und berufliches Engagement. Arbeitsengagement wiederum bündelt Vitalität, Hingabe und Aufnahmefähigkeit (Hakanen, Bakker & Schaufeli, 2006). Gleichzeitig birgt ein zu hohes Engagement zusammen mit überhöhten Erwartungshaltungen das Risiko an Burnout zu erkranken (vgl. Typ A des AVEM unten) – mit entsprechenden Folgen für die Unterrichtspraxis der Lehrpersonen (Schumacher et al., 2009). Schaufeli und Bakker weisen daher darauf hin, dass Arbeitsengagement nicht das positive Gegenstück zu Burnout darstellt, wenngleich es in vielen Studien mit derselben Skala gemessen worden sei. Auch könnten Personen ein geringes Arbeitsengagement aufweisen, ohne an Burnout erkrankt zu sein (Schaufeli & Bakker, 2003), (Schaufeli & Bakker, 2004). Der so genannte Schon-Typ in der Klassifikation von Schaarschmidt und Fischer ist charakteristisch derart angelegt (Schaarschmidt & Fischer, 1997). Dass eine geringe Arbeitszufriedenheit mehr bedeuten kann als nur eine Burnout-Erkrankung zeigt sich auch in der Abhängigkeit der Arbeitszufriedenheit von erlebter externer Kontrolle. Vor allem zeigt sich aber in Studien in diesem Kontext immer wieder eine sehr große Zahl befragter Lehrkräfte mit ihrer Tätigkeit zufrieden (Dick, 2006).

Burnout wiederum ist bei Lehrkräften durch das Gefühl verminderter Leistungsfähigkeit, Unzufriedenheit mit der (daraus resultierenden) beruflichen Leistung und einer teilweise zynischen Einstellungen gegenüber Schülerinnen und Schülern und Kollegen gekennzeichnet (Bromme & Rheinberg, 2006). Studien mit deutschen Lehrkräften über das Ausmaß von Burnout-Erkrankungen sprechen von einer Verbreitung von 10 bis 30% erkrankten Lehrkräften (Harazd, Gieske & Rolff, 2009). Rauin zeigte in einer längsschnittlich angelegten Studie, dass bei zehn Prozent der Berufsanfänger zumindest schon eine Überforderung vorhanden ist. Allerdings handelt es sich bei diesem Teil vorwiegend um Personen, die schon im Studium entsprechende Tendenzen zeigten (Rauin, 2007).

Belastung und Beanspruchung

Klusmann stellt in der Einführung zu den vier Studien zum Zusammenhang von Lehrerbelastung und Unterrichtsqualität auch für diesen Bereich einen Begriffswirrwarr fest. Begriffe wie Belastung, Beanspruchung, Stress und Burnout seien häufig nicht trennscharf, als Oberkategorie für verschiedene psychische Reaktionen auf berufliche Anforderungen verwendet worden. Gemeinsam sei den divergierenden Konzepten aber ihre Relevanz zum einen für die individuelle psychische Gesundheit und zum anderen für die Arbeitsleistung (Klusmann, 2008a). Nach Böhm-Kasper (2004), dessen Modell diese Arbeit an dieser Stelle folgt, sind die Begriffe Belastung und Beanspruchung folgendermaßen voneinander zu

treffen: Unter Belastung wird die Gesamtheit der erfassbaren Einflüsse (z.B. spezifische Arbeitsbedingungen) verstanden, die von außen auf die Lehrkraft einwirken. Diese können als Herausforderung oder Last wahrgenommen werden. Mit Beanspruchung werden hingegen die Auswirkungen dieser Belastungen bezeichnet. Beanspruchung stellt folglich als erlebte Belastung eine individuelle Größe dar. Die Belastung beinhaltet beispielsweise die Möglichkeit der Kooperation innerhalb der Schule als Qualität der Sozialbeziehungen und Qualität der Sachbeziehungen. Zusammen ergeben diese das situative Bedingungsfeld. Einflussgrößen wie die Arbeitszeit als objektive Anforderungen und das Alter oder andere Persönlichkeitsmerkmale als individuelle Voraussetzungen bilden zusammen das situationsübergreifende Bedingungsfeld im Modell von Böhm-Kasper (Fussangel, Ditzinger, Böhm-Kasper & Gräsel, 2010).

Wie das Verhältnis von Belastung und Beanspruchung zu verstehen ist, lässt sich an der Lehrerkooperation verdeutlichen. Die Kooperation innerhalb der Schule kann erlebt als soziale Unterstützung Belastungen reduziert wahrzunehmen und beispielsweise Burnout-Erkrankungen vorbeugen.[79] Dies lässt sich derart deuten, dass die Kooperation als Ressource für Hilfe bei Problemen und Bestätigung für das eigene Tun angesehen wird. Die Kooperation kann aber auch zu einer höheren Beanspruchung führen, wenn Kooperation als Selbstzweck, als erzwungene Zusammenarbeit angesehen wird und die Bedingungen für Kooperation in der Schule ungünstig sind (Fussangel et al., 2010). Kooperation bedeutet darüber hinaus auch einen Verzicht auf einen Teil der Autonomie, die bei vielen Lehrkräften zum Wert ihrer Arbeit zählt, und wird daher eher als Beanspruchung gesehen. Haradz und Gieske fand in ihrer Untersuchung zur Rolle der Schulleitung für die Lehrergesundheit allerdings, dass drei von vier Lehrern das Arbeitsklima im Kollegium eher als entlastend bezeichneten (2009), ähnlich in den Studien von van Dick (2006).

Nach van Dick werden vor allem folgende Faktoren von Lehrkräften als belastend empfunden: (1) zu große Klassen, (2) Probleme mit Schülerinnen und Schülern, (3) administrative Probleme, (4) Probleme mit Kollegen, (5) Probleme mit Erziehungsberechtigten und (6) fehlende Anerkennung durch die Öffentlichkeit. (Dick, 2006). Haradz und Gieske legten Lehrern 22 Items vor, die diese bzgl. der empfundenen Belastung bewerten sollten. Hier wurden besonders die Items als belastend eingestuft, die Haradz und Gieske als Reform- und Verwaltungsarbeit klassifizieren (Einführung der Kopfnoten, administrative Pflichten, Einführung zentraler Prüfungen, Gremien- und Konferenzarbeit) (Harazd et al., 2009).

Commitment, Pflichtbewusstsein und Distanzierungsfähigkeit

Grundsätzlich wird von Lehrkräften in besonderer Weise erwartet, dass sie sich mit der komplexen Aufgabe ihres Berufs stark identifizieren. Die Bindung von Zielen an das Selbstkonzept wird als Commitment bezeichnet (Spörl, 2009). Partizipation der Mitarbeiter

[79] Keller-Schneider (2010) konnte hingegen keinen Effekt von sozialer Unterstützung, also auch nicht durch Kooperation im Kollegium auf die erlebte Beanspruchung feststellen.

bei der Zielsetzung oder Begründungen für die Relevanz der Ziele durch die leitenden Personen erhöhen die Zielbindung. Eine große Bedeutung kommt aber auch der erlebte Gerechtigkeit zu (Nerdinger et al., 2008). Speziell für Lehrkräfte sollte es folglich eine Rolle spielen, ob ihre speziellen Bedingungen (Klassenzusammensetzung) angemessen berücksichtigt werden, wenn sie sich Leistungszielen (Lernzielen, Durchführung der Lernstandserhebungen) unterordnen sollen.

Aber es kann auch ein Zuviel an Bindung geben. In dem Fall spricht man von Overcommitment. Gefährlich wird es auch, wenn Lehrkräfte glauben, die von Bildungspolitik, -administration und -forschung aufgestellten Anforderungen und Standards an das professionelle Handeln von Lehrern und Lehrerinnen als einzelne Person erfüllen zu müssen. Daher ist es für Lehrkräfte ebenfalls äußerst relevant zu wissen, dass ihr Handeln nicht allein Resultat ihrer eigenen Motivation ist, sondern im Kontext des Arbeitsumfelds gesehen werden muss. Viele Aufgaben sind in Bezug auf die einzelne Lehrkraft formuliert, können aber nur kollektiv innerhalb des Kollegiums und in Zusammenarbeit mit den jeweiligen Erziehungsberechtigten erfüllt werden (Schumacher et al., 2009).

Den Grad der richtigen Bindung an die Ziele von Schule zu finden, stellt eine wichtige Aufgabe dar. An dieser Stelle kommt die Distanzierungsfähigkeit ins Spiel, die es bei hoher Ausprägung erlaubt, zu hohen Anforderungen durch eine zeitliche und räumliche Begrenzung zu begegnen. Die nötige Distanz zu wahren, wird allerdings durch die für die Schule typischen Gedanken und Gefühle, die aus zwischenmenschlichen Beziehungen entstehen, besonders erschwert (Schaarschmidt, 2009). Zusätzlich ist auch die Möglichkeit zur räumlichen und zeitlichen Abgrenzung von beruflichem und privatem Bereich aufgrund fehlender materieller Ressourcen am Arbeitsplatz zur Vor- und Nachbereitung und nur am Abend für Gespräche erreichbarer Erziehungsberechtigte gefährdet. Auch am Arbeitsplatz selbst sind – selbst in formalen Pausen – fast keine Möglichkeiten zur Entspannung und kurzfristigen Regeneration gegeben. Schaarschmidt (2009) stellt fest, dass die hohen Belastungen der Lehrkräfte im Vergleich zu anderen Berufsgruppen zu wachsenden Defiziten in den Widerstandsressourcen (vor allem eine Abnahme der Distanzierungsfähigkeit und der inneren Ruhe) führen. Durch den hohen Anteil von selbstgestalteter Arbeitszeit - Lehr, Hiller und Keller veranschlagen dafür mit Rückgriff auf Dorsemagen und Kollegen bis zu fünfzig Prozent (Lehr et al., 2009) - sind Lehrkräfte zusätzlich gefährdet.

Commitment ist eng verwandt mit dem allgemeinen Persönlichkeitsfaktor Gewissenhaftigkeit und wird auch als Zielstrebigkeit und Arbeitswille bezeichnet und stellt quasi eine auf den Beruf ausgerichtete Ausprägung davon dar (Barrick & Mount, 1991; McCrae & Costa, 2006; Saum-Aldehoff, 2007). Die Gewissenhaftigkeit expliziert die Sorgfältigkeit einer Person im Umgang mit ihr auferlegten Aufgaben. Eine hohe Ausprägung der Gewissenhaftigkeit wird mit Personen verbunden, die sehr sorgfältig sind, sich stets an Vereinbarungen halten und häufiger zu Perfektionismus neigen (Müthing, 2005). Gewissenhaftigkeit gilt als eine derjenigen Persönlichkeitsfaktoren, die in der Schnittmenge der meisten drei- bis sechsfaktoriellen Personeneigenschaftsfaktorenmodelle liegen und

kulturübergreifend repliziert werden konnten (Asendorpf, 2007). Anders als das Zusammenspiel von Commitment und Distanzierungsfähigkeit stellen Eigenschaftskonstrukte keine ausbaubare Fähigkeit dar, sondern sind als Traits konzipiert. Die Grundidee hinter Persönlichkeitsmodellen ist eine sparsame Beschreibung von Personen und Reduzierung von wahrscheinlichen Handlungen auf Persönlichkeitseigenschaften. Entsprechende Messinventare wie der NEO-Persönlichkeitsinventar (NEO-PI-R) von Costa und McCrae (1992, 2006) oder der Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP) nach (Hossiep & Paschen, 2003) sollen es erlauben, das Verhalten von Menschen weitestgehend aufgrund ihrer Eigenschaften vorherzusagen und werden daher häufig in der Personalauswahl eingesetzt.

Barrick und Mount verglichen in einer Meta-Analyse die Vorgesetztenbeurteilungen in fünf Berufsgruppen mit den Klassifikationen der Big Five von Costa und McCrae. Dabei stellte sich der Faktor Gewissenhaftigkeit als bester Vorhersagefaktor für die Vorgesetztenbeurteilung über Eignung für den Arbeitsplatz und Nutzen von Fortbildungsmaßnahmen heraus (Barrick & Mount, 1991). Auch Saum-Aldehoff berichtet ähnliche Ergebnisse aus verschiedenen Studien (Saum-Aldehoff, 2007). Kröner, Sparfeldt, Buch, Zeinz und Rost haben Leistungsangst und Zusammenhänge mit den Big Five untersucht. An einer Stichprobe aus 296 Studierenden des Haupt- und Realschullehramts konnten sie zeigen, dass die Sub-Konstrukte „Gefahrenkontrolle durch produktives Arbeiten“ leicht positiv und „Situationskontrolle durch Vermeiden und Mogeln“ leicht negativ mit Gewissenhaftigkeit (r.25 bzw. r-.31) korrelieren. Sie erklären dies, indem sie gewissenhafteren Personen eine bessere Vorbereitung auf Prüfungen unterstellen, die Vermeiden und Mogeln überflüssig werden lässt (Kröner, Sparfeldt, Buch, Zeinz & Rost). Trotz dieses Relationsbefundes ist ähnlich dem Overcommitment auch ein Zuviel an Gewissenhaftigkeit möglich, welches die ebenfalls negativen Folgen von überhöhtem Perfektionsstreben nach sich zieht (Saum-Aldehoff, 2007).

Berufliche Aspekte der Lehrperson und Unterrichtsqualität

In den vorherigen Erläuterungen und Befunden zu den personenbezogenen Kognitionen im Kontext des beruflichen Beanspruchungserlebens lässt sich bereits der Dualismus von Arbeitsleistung als Verbindung von Leistungsbereitschaft und Leistungsfähigkeit erkennen. Berufliche Leistung verbindet den Arbeitenden mit dem Produkt seiner Arbeit. Um dies konkreter zu erfassen, benötigt es passende Modelle. Das erweiterte Job-Demand-Resourcen-Model (JD-R-Modell) (Demerouti, Bakker, Nachreimer & Schaufeli, 2001; Prieto, Soria, Martínez & Schaufeli, 2008; Schaufeli & Bakker, 2004), und die Arbeitsbezogene Verhaltens- und Erlebensmuster (AVEM) (Schaarschmidt & Fischer, 1997, 2008) ermöglichen diese Verknüpfung von unterschiedlichen Standpunkten aus. Das JD-R-Modell betrachtet das Verhältnis von beruflichen Anforderungen der Organisation und (externen) beruflichen Ressourcen des Arbeitenden. Das AVEM klassifiziert Arbeitende aufgrund persönlicher beruflicher Ressourcen. Zuerst wird das JD-R-Model erläutert und der Leistungsaspekt

diskutiert, anschließend wird der AVEM-Ansatz dargestellt und es werden wichtige Befunde zur Beanspruchung von Lehrkräften und deren Wirkung auf den Unterricht dargestellt.

Das erweiterte Job Demand-Resources-Model

Wie Klusmann richtig herausstellt, geschieht insbesondere trotz der relativ großen Autonomie in der Arbeitsgestaltung bei Lehrkräften die Berufsausübung selbstverständlich nicht unabhängig vom schulischen Kontext. Vor allem sind auch im Berufsfeld Schule Arbeitsziele durch fremd aufgestellte Anforderungen gesetzt. Klusmann schlägt deswegen vor, ein organisationspsychologisches Modell zu verwenden, das auf Demerouti, Bakker, Nachreiner und Schaufeli zurückgeht und in den letzten Jahren von Bakker und Schafeli um die Differenzierung zwischen zwei Entwicklungssträngen bzw. von Prieto und Kollegen um personale Ressourcen erweitert wurde (Demerouti et al., 2001; Prieto et al., 2008; Schaufeli & Bakker, 2004): das Job Demands-Resources Model (JD-R-Modell). Das Modell fokussiert ursprünglich auf (externe) Ressourcen und Anforderungen im beruflichen Kontext und dient der Modellierung von Burnout-Prozessen über die Differenz von Ressourcen und Anforderungen im Beruf (Demerouti et al., 2001). Unter beruflichen Anforderungen (Job Demands) werden solche physischen, psychischen, sozialen und organisatorischen Aspekte der Arbeit verstanden, die permanente physische und mentale Anstrengung verlangen und dementsprechend physische und psychische Kosten verursachen. Ressourcen des Arbeitskontextes (Job Resources) sind physische, psychische, soziale und organisatorische Aspekte der Arbeit, die einen der drei folgenden Erträge begünstigen: a) Erreichen bestimmter Arbeitsziele ermöglichen, b) Reduzieren beruflicher Anforderungen, die Reduzierung verursachter Kosten oder c) Förderung der persönlichen Entwicklung und des persönlichen Wachstums (Demerouti et al., 2001). Die Erträge a) und c) werden im modifizierten Modell zu Arbeitsleistung zusammengefasst.

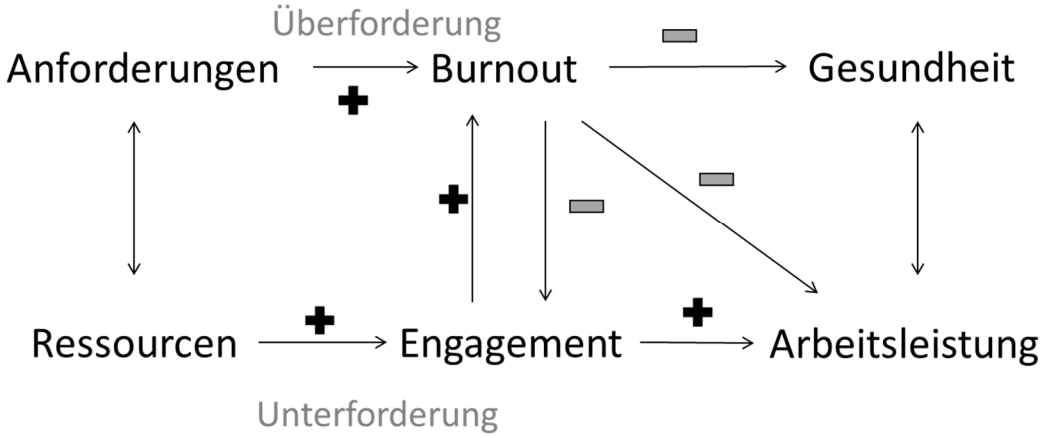


Abbildung 4.2 Das Job-Demand-Resource-Modell nach Schaufeli & Bakker (Übersetzung und Modifikation durch den Verfass.)

Problematisch an dem ursprünglichen JD-R-Modell ist der Fokus auf negative Entwicklungen. Eine gleichzeitig aktive Rolle des Berufstätigen ist in diesem Modell nicht vorgesehen. Dadurch bietet das Modell keine Erklärungskraft für Folgen einer positiven Differenz von Job Resorces und Job Demands. Job Demands sind in diesem Modell nicht als Chancen modellierbar. Entsprechend kann eine vorhandene Partizipation auch nur genutzt werden, um Anforderungen zu reduzieren, aber nicht um übererfüllte Anforderungen als Anlass zu nehmen, die Anforderungen dementsprechend anzupassen. Dem Modell nach führt ein Überschuss von Ressourcen im Verhältnis zu den Anforderungen zu einem Überengagement, welches auch in einem Burnout enden kann.

Ressourcen können sowohl als intrinsischer Motivator als auch als extrinsischer Motivator verstanden werden. Intrinsischer Motivator sind Ressourcen, da sie als Wert für sich angesehen werden. Im JD-R-Modell sind (externer) Ressourcen Rückmeldungen, Bezahlung, Rechenschaftslegung, Partizipation, Arbeitsplatzsicherheit und Unterstützung der Kollegen und Vorgesetzten (Demerouti et al., 2001). Ressourcen können nach Hobfoll und Buchwald auch interne Ressourcen sein und als Objektressourcen (Kleidung, das eigene Auto und Haus usw.) und Bedingungsressourcen (Familienstand, Gesundheit, berufliche Position usw.) Statuscharakter haben bzw. als persönliche Ressourcen (Fähigkeiten und Eigenschaften) und Energieressourcen (Zeit, Geld, Wissen) der persönlichen Weiterentwicklung dienen. Extrinsischer Moderator sind Ressourcen, wenn sie dem Erreichen von (beruflichen) Zielen dienen (Hobfoll & Buchwald, 2004; Prieto et al., 2008; Schaufeli & Bakker, 2004).

Berufliche Anforderungen (konkret: Arbeitszeiten und Arbeitsumfang, Zeitdruck, ungünstige physische Bedingungen der Arbeit und Kundenkontakt) werden stets als Forderungen verstanden, die mittels der vorhandenen Ressourcen zu erfüllen sind. Das JD-R-Modell dient ursprünglich der Erklärung, welche Folgen ein Defizit von Ressourcen zu Anforderungen hat,

nämlich emotionale Aufzehrung und sinkende Leistungsfähigkeit als Komponenten des Oldenburger Burnout-Begriffs (Schaufeli & Bakker, 2004).

In der Erweiterung des JD-R-Modells beschreiben Schaufeli und Bakker entsprechend zwei verknüpfte Prozess: Der Energie-Prozess als prototypischer Ablauf der Burnout-Forschung besteht aus einem Ressourcen-Anforderungs-Defizit, aus dem Burnout-Symptome resultieren, und diese wiederum führen zu gesundheitlichen Problemen (vgl. oberen Weg in Abb. 4.2). Dabei treten Burnout-Symptome genau dann auf, wenn zu hohe Anforderungen langfristig existieren und nicht durch kurzzeitiges starkes Engagement oder reduzierte Leistung aufgefangen werden können. Der Motivations-Prozess stellt als Parallele zum Energie-Prozess erst einmal statt der Burnout-Symptome das Arbeitsengagement als Mediator in die Verbindung zwischen dem Ressourcen-Anforderungs-Defizit und der aus zu geringem Arbeitsengagement resultierenden Arbeitsplatzunzufriedenheit (vgl. unteren Weg in Abb. 4.2). Ein wenig abstrakter gesehen wird durch den Motivations-Prozess das Verhältnis von Job Resources, Arbeitsengagement und Leistungszielen allgemein erklärt.

Für die Verknüpfung von Leistungszielen der Organisation mit den Zielen der Arbeitnehmenden rekurriert das JD-R-Modell auf die Job-Characteristics-Theorie (JCT) von Hackman und Oldham (1980). Mit der JCT erklären Hackman und Oldham eigentlich die intrinsische Arbeitsmotivation. Motivierend ist Arbeit danach in dem Ausmaß, wie sie in den folgenden fünf Kernjobdimensionen hohe Werte aufweist: Skill Variety, Task Identity, Task Significance, Autonomy und Feedback (Hackman & Oldham, 1980). Dieser Gedankengang der JCT ist ebenfalls Bestandteil des Motivations-Prozesses (Schaufeli & Bakker, 2004). Dazu muss man unterstellen, dass die arbeitende Person entsprechend dem Grad ihrer (intrinsischen) Arbeitsmotivation sich mit den Zielen der Organisation identifiziert.

Durch die Modellierung dieser zwei Prozesse wird Arbeitsleistung ein zweifaktorielles Modell aus Leistungsbereitschaft und Leistungsfähigkeit. Schaufeli und Bakker sehen die beiden Prozesse als parallele Stränge, die durch die Beziehungen von Job Demands und Job Resources, Job Resources und Burnout, Burnout und Engagement und den Folgen für Berufstätigen und Organisation verknüpft sind (Schaufeli & Bakker, 2004). Das Modell kann aber auch als Darstellung der beiden Elemente von Arbeitsleistung verstanden werden. Organisationsprozesse können nur erfolgreich sein, wenn die beteiligten Personen das richtige Maß an notwendigen Ressourcen und Motivation besitzen.

Die speziellen Anforderungen (Job Demands) an Lehrkräfte sind beispielsweise ein hoher Arbeitszeitumfang, kognitive Herausforderungen, emotionale Herausforderungen verursacht durch Empathie, Bewältigung von Rollenkonflikten (Lehrkräfte sollen fördern und selektieren), Konkretisierung von Aufgaben (Prieto et al., 2008), Umgang mit schwierigen Schülerinnen und Schülern (Hakanen et al., 2006), besonders aber Vermittlung von Wissen und Methoden als Hauptziel des Unterrichts (Baumert & Kunter, 2006). Letzteres lässt sich nach Klusmann weiter differenzieren. Während des Unterrichtens müssen die eigenen Ziele mit den Interessen verschiedener Schüler und Schülerinnen abgestimmt werden (Multidimensionalität der Ereignisse) und auf Ereignisse muss unmittelbar reagiert werden

(Unmittelbarkeit der Geschehnisse), während diese Ereignisse nur sehr schwierige vorhersehbar sind (Unvorhersehbarkeit der Geschehnisse) und Handlungen stets unter den Augen der Schülerinnen und Schüler geschehen (die Öffentlichkeit des Handelns) (Klusmann, 2008a). Zu den wichtigsten Job Resources zählen Prieto und Kollegen die Autonomie der Lehrkräfte und die soziale Unterstützung durch Kollegen und Kolleginnen bzw. durch die Schulleitung (Prieto et al., 2008). Weiter kann die Arbeitsplatzsicherheit dazugezählt werden (Hakanen et al., 2006), wenngleich diese – wie bereits angemerkt – für Lehrkräfte weniger wichtig zu seien scheint. Während in anderen Berufsgruppe Faktoren wie Kündigung, Bonuszahlungen oder Beförderung als externe Ressourcen und Motivationselemente verfügbar sind, stellt sich die Arbeitssituation bei Lehrkräften anders da. Auch wenn ein nicht unerheblicher Teil der Lehrkräfte in Angestelltenverhältnissen beschäftigt ist und Kündigungen und nicht vollzogene Vertragsverlängerungen potenzielle Faktoren dieser Art sein könnten, verfügen die meisten Lehrkräfte über die Ressource Arbeitsplatzsicherheit.[80] Rückmeldungen sind bei Lehrkräften sehr geachtet, fehlen aber häufig (s. Abschnitt Belastung und Beanspruchung). Auch durch die Umstellung von Input- zu Outputsteuerung haben Lehrkräfte Freiheiten erhalten, die als Partizipationsressource verstanden werden können.[81]

Der Begriff der Lehrerleistung

Im Abschnitt Arbeitszufriedenheit und Arbeitsengagement wurde neben der Arbeitszufriedenheit als Wert an sich für den Arbeitenden auch die Mittel-Funktion der hier skizzierten personenbezogenen Kognitionen angesprochen. Wird Arbeitszufriedenheit über das Arbeitsengagement als Moderatorvariable für Leistung angesehen, sind Leistungsunterschiede das eigentlich Ziel der psychologischen Forschung. Grundsätzlich ist der Leistungsbegriff doppeldeutig. Leistung als Begriff beinhaltet als Aspekt sowohl das Ergebnis eines Verhaltens als auch einen Handlungsaspekt. Unter den Leistungsbegriff wird dabei dasjenige Verhalten gefasst, welches für die Ziele einer Organisation relevant ist. Leistung als Handeln meint selbstverständlich nicht, irgendeine Handlung auszuführen, sondern wird durch evaluative Prozesse bestimmt (Sonnentag & Frese, 2002). Während bei einem produzierenden Industrieunternehmen die Leistung der Mitarbeiter in Abhängigkeit zum Produkt relativ klar umfassend beschrieben werden kann, ist die Definition des Leistungsbegriffs für Lehrkräfte wesentlich komplizierter. Weder der Kontext, in dem Lehrkräfte eine Arbeitsleistung erbringen, noch das Maß zur Messung einer irgendwie umschriebenen Leistung sind vergleichbar eindeutig. Diesen Schwierigkeiten kann man nicht

[80] So können insgesamt externe Motivationselemente für Lehrkräfte nicht flächendeckend herangezogen werden. Dies liegt zu einem großen Teil am weiterhin stark verbreiteten Beamtenstatus der Lehrkräfte, aber auch an der bereits dargelegten Schwierigkeit, insbesondere Schülerlernleistungen als Indikatoren für Lehrerleistung heranzuziehen.

[81] Eine der wichtigsten und oft übersehenen Ressourcen von Lehrkräften ist der Ertrag der mehrjährigen Ausbildung und der praktischen Erfahrung. Da diese Ressource aber eine eigene Dimension im Lehrer-Handlungskompetenzmodell darstellt und bereits in einem eigenen Abschnitt behandelt wurde, wird an dieser Stelle nicht weiter darauf Bezug genommen.

damit ausreichend begegnen, dass Forschung zur Arbeitsleistung bei Lehrkräften nur die Unterrichtstätigkeit in den Blick nimmt. Auch die Förderung der Motivation und des Interesses, aber auch die persönliche Entwicklung, die auch in Schulen zu großen Teilen außerhalb der Unterrichtszeit gefördert wird, und gehören zu den Kernaufgaben von Lehrkräften (Kunter, 2005). Schließlich gehören zu einem mehrdimensionalen Leistungsbegriff auch die weiterentwickelten Arbeits- und Organisationsprozesse und -abläufe (Sonnentag & Frese, 2002). Die tatsächliche Arbeitsleistung erstreckt sich folglich über die Leistung innerhalb des Unterrichts hinaus und müsste in ein Leistungsmodell integriert werden (Tenorth, 2006). Die Leistung von Lehrkräften ist damit ein typisches Beispiel dafür, dass ein am Produkt orientierter Leistungsbegriff und ein auf das Handeln abzielender Leistungsbegriff zwar ähnlich, aber nicht vollständig überlappend sind. Gleichzeitig ist es äußerst schwierig, Leistung im Sinne des Handlungsaspekts zu beschreiben, ohne auf einen Ertrag Bezug zu nehmen (Sonnentag & Frese, 2002). Die Leistung von Lehrkräften kann folglich andersherum auch nicht losgelöst von Veränderungen im Verhalten und Können von Schülerinnen und Schülern gesehen werden.

Ein Leistungsbegriff, der sich auf den Aspekt des Unterrichtens beschränkt, muss zusätzlich noch eine andere Klippe umschiffen. Auch wenn man die Probleme beim Messen von Schülerleistungen außen vor lässt, wird eine Gleichsetzung der Lehrerleistung mit der Schülerleistung den Anforderungen, denen Lehrkräfte ausgesetzt sind, nicht gerecht. Schon zu Beginn dieses Kapitels ist auf die doppelte Unsicherheit in der Interaktion zwischen Lehrkräften und Schülerinnen und Schülern im Unterricht hingewiesen worden. Diese doppelte Unsicherheit ergibt sich auch für das Messen der Schülerleistung. Schülerlernleistungen sind stark von der Mitarbeit der Schüler und Schülerinnen selbst, ihren Vorkenntnissen, der elterlichen Unterstützung und anderen schulischen Erfahrungen abhängig (Tenorth, 2006). Entsprechend ist die Lehrleistung über die Lernleistung nur im jeweiligen Referenzrahmen zu beurteilen. Ein Ausweg aus diesem messpraktikablen Dilemma besteht darin, die Unterrichtsqualität als Prädiktor für die Lehrleistung zu verwenden. Die Gestaltung von Lerngelegenheiten ist dann das Maß für den Arbeitserfolg der Lehrkraft. Dies setzt allerdings Wissen darüber voraus, welche Qualitätsmerkmale guter Unterricht haben muss.[82] Befunde dieser Art sind somit Prozess-Produkt-Zusammenhänge (Klusmann, 2008a) und sind daher ebenfalls problematisch.

Die Lösung der aktuellen Forschung (beispielsweise im Projekt COACTIV) verknüpft die Perspektive des Lehrerexpertise-Ansatzes mit der erweiterten Differenzierung des Prozess-Produkt-Paradigmas in Sicht- und Tiefenstruktur. Elemente der Tiefenstruktur von Unterricht, die als lernförderlich eingestuft werden, werden mit Charakteristika eines Experten für das Lehren und Lernen verknüpft (Klusmann, 2008a). Dabei greift man auf die Beziehungsstrukturen zurück, wie sie bereits im Motivations-Prozess des JD-R-Models unterstellt wurde: beispielsweise auf die Job-Characteristics-Theorie von Hackman und Oldham (1980). Es wird unterstellt, dass ähnlich der fünf Kernjobdimensionen Skill Variety,

[82] Möglicherweise variiert auch das Qualitätsbündel in Abhängigkeit vom explizit vorhandenen Referenzrahmen.

Task Identity, Task Significance, Autonomy und Feedback der JCT die Aufgaben des Lehrerberufs ebenfalls grundsätzlich motivierenden Charakter besitzen und Lehrkräfte ihre Kernaufgabe – die Entwicklung von Schülerinnen und Schülern – als ihr Arbeitsziel übernehmen. Nach den gängigen aktuellen Theorien haben persönliches Beanspruchungserleben und situative äußerliche Belastungsfaktoren einen Einfluss auf die Unterrichtstätigkeit. Umgekehrt wirkt aber die eigene Überzeugung über die unterrichtliche Kompetenz auch auf das Wohlbefinden und bedingt sich schließlich gegenseitig (Klusmann, 2008).

Demnach wird weiter unterstellt, dass auch Lehrkräften prinzipiell bekannt ist, a) welche Qualität guter Unterricht aufweisen muss und Lehrkräfte diese Qualität erreichen wollen und b) Lehrkräfte wissen, welches Unterrichtshandeln für sie selbst das vorteilhafteste ist. In diesem Sinne gelingt ihnen im Idealfall eine Arbeitsleistung u.a. in Abhängigkeit ausreichender beruflicher Ressourcen.

Die empirische Befundlage bzgl. einer derart gefassten Lehrerleistung ist allerdings sehr dürftig. Einzig Klusmann konnte in zwei Studien einen Einfluss von persönlichen Merkmalen von Mathematiklehrkräfte auf das durch Schüler und Schülerinnen wahrgenommene Unterrichtsverhalten finden, ein Zusammenhang mit der Schülerleistung in Mathematik zeigte sich hingegen nicht (Klusmann, 2008).

Eine Typenklassifikation im Arbeitsleben

Unabhängig von konkreten Anforderungen, aber auf Grundlage der drei Bereiche Arbeitsengagement, Widerstandsfähigkeit und beruflich verbundenen Emotionen haben Schaarschmidt und Fischer (Schaarschmidt & Fischer, 1997, 2008) vier Bewältigungsmuster (Arbeitsbezogenes Verhaltens- und Erlebensmuster – AVEM) konzipiert, welche berufliche Beanspruchung voraussagen sollen und in verschiedenen Studien reproduziert werden konnten (Klusmann, 2008; Klusmann, Kunter & Trautwein, 2009; Schaarschmidt, 2009). Diese beruflichen Verhaltensstile lassen sich folgendermaßen klassifizieren: Der Gesundheitstyp (Typ G) weist ein hohes, aber nicht zu hohes Arbeitsengagement auf. Auch seine Widerstandsfähigkeit und seine auf seinen Beruf bezogenen Emotionen sind positiv. Menschen, die zu diesem Typ zählen, vereinen sowohl gute Voraussetzungen zur Bewältigung der Arbeit als auch eine hohe Profession der eigenen Person gegenüber in sich. Der Schontyp (Typ S) ist dem Typ G in den Bereichen Widerstandsfähigkeit und beruflich verbundenen Emotionen sehr ähnlich und besitzt nur ein geringes Risiko aufgrund seiner Arbeitsweise zu erkranken. Er ist aber nur sehr gering engagiert und daher nur sehr bedingt in der Lage, seine Aufgaben – insbesondere die an Lehrkräfte gestellten – zu erfüllen. Den Risikotypen (Typ A und Typ B) ist hingegen gemeinsam, dass sie nur eine geringe Widerstandsfähigkeit besitzen und ihrem Beruf eher negativ gegenüber eingestellt sind. Typ A unterscheidet sich von Typ B dadurch, dass Typ A ein übersteigertes Arbeitsengagement aufweist während Typ B dort die niedrigsten Wert alle vier Typen erreicht. Menschen des

Typs A neigen zur Selbstüberforderung. Wenngleich sie wegen ihres hohen Engagements häufig geschätzt werden, besitzen sie ein hohes Burnout-Risiko.

Schaarschmidt schätzt den Anteil derart gefährdeter Lehrkräfte auf 25% (Schaarschmidt, 2009). Dieses Stadium ist für Menschen des Typs B bereits erreicht. Letzte Kraftreserven werden dazu benötigt, „irgendwie zu überleben“ (Klusmann et al., 2009). Es muss allerdings festgehalten werden, dass nur ein geringer Teil der untersuchten Personen sich eindeutig einer dieser vier Typen zuordnen lässt. Viele sind Mischtypen, d.h. sie weisen weniger als 95% Zuordnungswahrscheinlichkeit auf (Schaarschmidt, 2009).

Ein Vergleich mit anderen Berufsgruppen (Beamte des Strafvollzugs, der Polizei und der Feuerwehr sowie Pflegepersonal und Existenzgründer) zeigt für Lehrkräfte eine Typverteilung, mit der geringsten G-Typ-Gruppe (ca. 17%) und knapp 60% für die beiden Risikogruppen Typ A und Typ B. Dabei kann für alle untersuchten Berufsgruppen ähnliche psychosoziale Beanspruchung unterstellt werden (Schaarschmidt, 2005).

Befunde zum Arbeitsbezogenen Verhaltens- und Erlebensmustern bei Lehrkräften

Klusmann u.a. fanden in einer Längsschnittstudie (im Rahmen von COACTIV) mit einem zweiten Messzeitpunkt nach einem Jahr bei Mathematik-Lehrkräften negativere Wert in den Variablen „Emotionale Erschöpfung“, „Berufliches Erfolgserleben“ und „Lebenszufriedenheit“ bei Typ-A- und Typ-B-Lehrkräften und zumindest geringere Werte beruflichen Erfolgserleben bei Typ-S-Lehrkräften im Vergleich zu Lehrkräften, die eher dem Typ G zuzuordnen sind, wenn eine Entwicklungsvorhersage aufgrund der Typen-Zuordnung. getroffen wird. Daraus folgern sie, dass für Lehrkräfte der Risiko-Typen eine ungünstigere Entwicklung des Beanspruchungserlebens zu erwarten ist (Klusmann et al., 2009).

Die duale Struktur von Arbeitsleistung erlaubt es, diese mit der Konzeption des AVEM von zu identifizieren. Leistungsbereitschaft drückt sich im Konzept von Fischer und Schaarschmidt als Arbeitsengagement aus. Zur Leistungsfähigkeit zählt die Widerstandsfähigkeit (Klusmann, 2008a). Klusmann, Kunter, Trautwein und Baumert vermuten diesbzgl. mit Rückgriff auf das Vier-Typen-Modell des AVEM von Schaarschmidt und Fischer, „dass Lehrkräfte mit geringem Arbeitsengagement (Typ S, Typ B) und/oder geringen Fähigkeiten auf den Dimensionen Widerstandsfähigkeit (Typ A, Typ B), Resignationstendenz (Typ A, Typ B), Problembewältigung (Typ B) und innerer Ausgeglichenheit (Typ A, Typ B) den multiplen Anforderungen zur Gestaltung eines qualitätsvollen Unterrichts kaum gerecht werden können“ (Klusmann, Kunter, Trautwein & Baumert, 2006, S.5). Betrachtet man die absolute Verteilung der Lehrkräfte in verschiedenen Studien auf die vier Typen, erscheint diese Einschätzung recht pessimistisch. In keiner Studie konnten auch nur ein Drittel der Lehrkräfte dem G-Typ zugeordnet werden (Klusmann et al., 2006; Schaarschmidt & Fischer, 1997; Schaarschmidt et al., 1999; Schaarschmidt, 2005).

Maslach und Leiter folgern aus ihren Studien als Folgen von Burnout-Erkrankungen Defizite im Sozialverhalten gegenüber den Schülerinnen und Schülern und in einer geringeren Gründlichkeit bei der Unterrichtsvorbereitung (Maslach & Leiter, 2006). Gründe für die Wahl repetitiver Unterrichtsformen – welches schon von Diedrich und Kollegen als Merkmal der Unterrichtsqualität untersucht wurde – könnten mangelnde Management-Fähigkeiten, erhöhte Ängstlichkeit aufgrund emotionaler Belastung, aber auch die Vorstellung einer an Disziplin und Ruhe orientierten Idealform effektiver Klassenführung sein (Klusmann et al., 2006).

4.2.4 Weitere personenbezogene Überzeugungen: Kompetenz- und Kontrollüberzeugungen von Lehrkräften

Überzeugungen über das eigene Selbst stellen die dritte Komponente der Erklärung von Handlungsabsichten dar. Ziele, Wertvorstellungen, Kausalitätsüberzeugungen und personenbezogene Kognitionen, wie sie hier in Form von Arbeitsengagement, Widerstandsfähigkeit und erlebter Belastung behandelt wurden, können Handlungsabsichten von Lehrkräften nicht allein erklären. Es braucht darüber hinaus auch Kontrollerwartungen und Kompetenzüberzeugungen. Für sich betrachtet haben Kontroll- und Kompetenzüberzeugungen sogar eine größere Aufklärungskraft als personenbezogene Kognitionen (Dick, 2006). Unter Kontrollüberzeugungen werden Kognitionen verstanden, bei denen innere und äußere Einflüsse auf Handlungen, Handlungsresultate und Ereignisse erwartet werden. Kompetenzüberzeugungen sind über verschiedene Situationen hinweg generalisierte und auf neue Situationen übertragende Erwartungen bzgl. der Aufnahme, Aufrechterhaltung und Steuerung zielgerichteter Handlungen (Preiser, 2006). Es wird angenommen, dass die personenbezogenen Kognitionen zwar die Interpretation von Ereignissen und dadurch das daraus resultierende Handlungsabsichten und die resultierenden Emotionen beeinflussen, aber immer erst im Zusammenspiel mit den Kontrollerwartungen und Kompetenzüberzeugungen wirksam werden. Personenbezogene Kognitionen, Kontroll- und Kompetenzüberzeugungen fungieren dann insgesamt als Moderatorvariablen für das Handeln, indem sie die Beziehung zwischen Zielen und deren Bewertung einerseits und dem Ausmaß persönlicher zielorientierter Aktivität andererseits modifizieren (Hofmann & Preiser, 1989). Behandelt werden diese Kontroll- und Kompetenzüberzeugungen in der Forschung unter der Kategorie Selbstkonzept.

Das Selbstkonzept

Das Selbstkonzept insgesamt kann als eines der wichtigsten Konstrukte in den Sozialwissenschaften bezeichnet werden (Marsh, 2005). Unzählige Arbeiten beschäftigen sich mit diesem Konstrukt, welches aus dem Selbst herausgearbeitet wurde und auf

Überlegungen von William James zurückgeht. Ein erster Schritt ist die Unterscheidung von Selbstkonzept und Selbstwertgefühl als kognitiver und affektiv-evaluativer Komponente des Selbst. Das Selbstkonzept wird weiterhin in Subkonzepte wie „Selbstwirksamkeitserwartung“ (SWE), „schulisches Selbstkonzept“ oder „mathematik-didaktisches Fähigkeitsselbstkonzept“ differenziert, welche sich weiter spezifizieren lassen wie beispielsweise „mathematische Fähigkeitsselbstkonzept“ und „verbale Fähigkeitsselbstkonzept“. Uneins ist sich die Forschergemeinde bei der Frage, ob das Selbstkonzept eine hierarchische Struktur hat, wie Hattie und Marsh sie vertreten (Hattie, 1992), (Hattie & Marsh, 1996), oder beispielsweise netzwerkartig organisiert ist, wie andere Forscher dies vertreten (Moschner & Dickhaus, 2006).

Die Sprechweise von „dem Selbstkonzept“ einer Person wird der tatsächlichen Verwendung in der psychologischen Forschung allerdings wenig gerecht. Wenn Bong und Clark davon sprechen, dass Selbstwirksamkeitserwartung und Selbstkonzept zwei unterschiedliche (aber hierarchisch ähnliche) Konstrukte seien (Bong & Clark, 1999), ist dies nur insoweit sinnvoll, als dass Selbstkonzept in einer speziellen Form verstanden wird, wie es beim Fähigkeitsselbstkonzept oder dem mathematischen Fähigkeitsselbstkonzept geschieht. Richtiger differenziert ist die Selbstwirksamkeitserwartung Teil des Selbstkonzepts, denn die SWE ist unbestreitbar Teil dessen, was eine Person über sich denkt, und dies entspricht dem von James konzipiert Selbst. Mit Fähigkeitsselbstkonzept ist die Vorstellung einer Person gemeint, welche Fähigkeit sie sich selbst in welchem Ausmaß zuschreibt. Zusätzlich scheint es sinnvoll, auch die Selbstdarstellungsstrategien diese Fähigkeit betreffend unter den Begriff zu denken, da diese in der Erhebung nicht getrennt werden können (Lipowsky, 2003).

Wichtige Forschungsfelder im Bereich der Selbstkonzeptstudien sind die Frage nach dem Zusammenhang von Feedback und Fähigkeitsselbstkonzept und nach dem Zusammenhang von Leistungen und Fähigkeitsselbstkonzept. Während die Wirkung der Leistung und darauf folgender Rückmeldungen auf das Fähigkeitsselbstkonzept recht gut erforscht und bestätigt ist, ist es ungleich schwieriger Belege zu finden, die einen Einfluss des Fähigkeitsselbstkonzepts auf die Leistung nachweisen. Trotzdem scheint ein gewisser Einfluss wahrscheinlicher als eine völlige Unabhängigkeit, wenn man Studien unter Schülerinnen und Schülern heranzieht (Dickhäuser, 2006). Das dritte Forschungsfeld zum Selbstkonzept im Schulkontext ist das berufliche Selbstkonzept von Lehrkräften (berufliches Selbstverständnis, Berufszufriedenheit, Berufswahlmotivation) (Bennewitz, 2009; Bönsch, 2008; Kiel, Geider & Jünger, 2004).

Ungeklärt ist nach aktuellem Forschungsstand auch, ob das Fähigkeitsselbstkonzept als relativ zeitlich stabil oder variabel anzusehen ist. Es wird vermehrt davon ausgegangen, das Fähigkeitsselbstkonzept sei umso variabler desto spezifischer und damit weniger subjektiv bedeutsamer es sei (Moschner & Dickhaus, 2006). Das Fähigkeitsselbstkonzept entwickelt sich durch die Wahrnehmung von gelösten oder nicht gelösten Aufgaben, durch soziale Vergleiche oder auch durch verbale Äußerungen anderer. Anders als bei der SWE, welche sich im Ausdruck immer auf das Erreichen von (gesetzten) Zielen bezieht, können

grundsätzlich nach Schön et al. vier Bezugsnormen unterschieden werden, mit denen das Fähigkeitsselbstkonzept ausgedrückt wird: individuell, sozial, kriterial und absolut (Schöne, Dickhäuser, Spinath & Stiensmeier-Pelster, 2002). Beispiele hierfür sind: „Ich bin nun besser in Mathematik als ich es noch vor drei Jahren war." „Ich bin in meiner Klasse einer der besten in Mathematik." „Wenn ich mir meine Mathematikklausuren ansehe, denke ich, dass ich das ganz gut kann." „Ich bin gut in Mathematik." Wird das Fähigkeitsselbstkonzept gebildet, stehen also alltägliche, durchschnittliche Maßstäbe im Vordergrund.

Zusätzlich spielen Vergleiche zwischen einzelnen Domänen eine Rolle. Wer durchschnittlich in Mathematik ist und zurückgemeldet bekommt, in Deutsch sehr gut zu sein, wird in Mathematik trotz durchschnittlicher Leistungen ein niedriges Fähigkeitsselbstkonzept entwickeln (Moschner & Dickhaus, 2006). Da das Fähigkeitsselbstkonzept Teil des Selbstbildes ist, fällt es unter den Bezugsrahmen der von Festinger postulierten Theorie des sozialen Vergleichs (Festinger, 1954) und der von ihm aufgestellten Theorie der kognitiven Dissonanz (Festinger, 1957).

Die Selbstwirksamkeitserwartung

Eine andere Facette des Selbstkonzepts sind Selbstwirksamkeitserwartungen. Umstritten ist, ob es sich dabei um Kompetenz- oder bzw. und Kontrollüberzeugungen handelt. Lipowsky beschreibt in seiner Dissertation in dem entsprechenden Abschnitt zu Kontroll- und Kompetenzüberzeugungen die SWE und allgemeine Formen des Selbstkonzepts als Kompetenzüberzeugung, nennt aber keine zusätzliche Kontrollüberzeugung (Lipowsky, 2003). Die Verortung der Selbstwirksamkeit als Kompetenzüberzeugung oder Kontrollüberzeugung gestaltete sich forschungshistorisch betrachtet auch als schwierig. Rotter hat die Lehrer-SWE als Kontrollüberzeugung konzipiert und mit folgenden Items gemessen:

1) When it comes right down to it, a teacher really can't do much because most of a student's motivation and performance depends on his or her home environment.

2) If I try really hard, I can get through to even the most difficult or unmotivated students. (nach Schulte, 2008).

Eine ähnliche Einordnung nehmen Hofmann und Preiser 1989 vor, die die SWE als „spezifische Kontrollüberzeugung" bezeichnen (Hofmann & Preiser, 1989).

Zur Kontrollüberzeugung wird neben der Selbstregulation aber auch eine Form der Autonomie gezählt, welche von den meisten SWE-Konzepten nicht abgedeckt wird (Krapp & Ryan, 2002). Dementsprechend unterscheiden Köller und Möller mit Verweis auf Skinner zwischen der Überzeugung, über entsprechende Mittel zu verfügen („agency beliefs" – entspricht SWE), und der Kontrollüberzeugung („control beliefs") (Köller & Möller, 2006).

Auch in dieser Arbeit wird die Selbstwirksamkeitserwartung in Anlehnung an Bandura als motivational-orientiertes Element der Kompetenz- und Selbstregulationsüberzeugung aufgefasst und ist als solche in das Selbstkonzepts integriert. Sie ist insofern eine Kontrollüberzeugung, wie sie die Überzeugung umfasst, die eigenen Fähigkeiten als Lehrkraft im Unterricht einsetzen zu können. Dazu müssen zwei Fragen reflektiert werden: die Frage nach dem Einsatz unterrichtlicher Kompetenz („personale Kontrolle“) und die Frage nach der Beeinflussung der Schüler und Schülerinnen („externe Kontrolle“). Anders als das Konzept der SWE bei Bandura ursprünglich vorsieht, handelt es sich bei dieser speziellen Form der SWE nicht nur um eine Kompetenzüberzeugung, sondern sie kann auch als Kontrollüberzeugung angesehen werden (vgl. Beispiel bei Preiser, 2006).

Das Konzept der Selbstwirksamkeitserwartung beschreibt die Überzeugung einer Person, auch in zukünftigen, schwierigen Situationen mittels adaptiver Handlungsmöglichkeiten selbst gesteuert erfolgreich zu sein. Es kann als ein spezielles Konzept aus der Klasse der Optimismus-Konzepte angesehen werden. Optimistische Menschen attribuieren positive Ereignisse internal und negative external, weniger optimistische Menschen betreiben genau umgekehrte Attributionen (Schwarzer & Jerusalem, 2002). Besondere Beachtung muss beim Konzept der SWE „schwierigen Situationen“ geschenkt werden, denn das Konzept der Selbstwirksamkeitserwartung fokussiert auf Belastungsspitzen, welches sich auch in den Itemformulierungen entsprechender Skalen niederschlägt („Die Lösung schwieriger Probleme gelingt mir immer, wenn ich mich darum bemühe.“). Bei der Selbstwirksamkeitserwartung wird folglich etwas anderes gemessen als beim Fähigkeitsselbstkonzept, welches u.a. allgemeiner, aber referenzgruppenabhängig scheint. Wenngleich Fähigkeitsselbstkonzept und Selbstwirksamkeit zwei verschiedene Konstrukte darstellen (Bong & Clark, 1999), sind gewisse Parallelen zwischen beiden Konstrukten vorhanden. Auch empirisch lässt sich ein mittlerer Zusammenhang zwischen beiden Konstrukten finden, beispielsweise erhielten Pajares und Miller (1994 nach Bong & Clark, 1999) eine Korrelation von r.54 für das mathematisch-schulische Fähigkeitsselbstkonzept und die mathematische Selbstwirksamkeit, ähnlich Skaalvik und Rankin (1995, Bong & Clark, 1999).

Der Grad der Selbstwirksamkeitserwartung steuert, welche Ziele sich eine Person setzt und welche Handlungen sie dazu ausführt, aber auch, welche Anstrengung sie bereit ist für dieses Ziel zu investieren und mir welcher Ausdauer sie dies tut. Die Selbstwirksamkeitserwartung besitzt folglich eine volitionale und eine motivationale Komponente (Bandura 1997 nach Lipowsky, 2003). Personen mit höherer SWE setzen sich auch höhere Ziele. Unter „Ziel“ können sowohl Leistungsziele als auch Ziele anderer Lebensbereiche verstanden werden (Schwarzer & Jerusalem, 2002).

SWE kann aus verschiedenen Blickwinkeln bzw. zu verschiedenen Prozessstadien wirken: die Überzeugung zielführend zu agieren und entsprechende Arbeits- und Zeitmanagementstrategien zu nutzen („action self-efficacy“), Widerständen wie Ablenkung

zu widerstehen („resistance self-efficacy“), Rückschläge auszuhalten („coping self-efficacy“) und sich erholen zu können („recovery self-efficacy“) (Schwarzer & Jerusalem, 2002).

Die Selbstwirksamkeitserwartung einer Person steigt oder fällt, wenn (selbst gesetzte) Ziele durch eigenes Zutun erreicht oder verfehlt werden und dies gleichzeitig internal-stabil attribuiert wird. Sie generiert sich aber auch aus Äußerungen anderer („du schaffst das schon“) und durch Vergleich mit anderen Personen („was der kann, kann ich schon lange“). Dabei unterscheiden sich nach Tschannen-Moran und Hoy Berufseinsteiger von erfahrenen Lehrkräften. Ersterer entwickeln ihre SWE auch weiter, indem sie auf verbale Überzeugungen und Ressourcen der Schule zurückgreifen, während die berufserfahrenen Lehrkräfte ihre SWE auf vorangegangene Erfolge zurückführen (nach (Warner & Schwarzer, 2009). Viertens wird die Entwicklung einer höheren SWE unterstellt, wenn sich diese auf einen Bereich bezieht, der zum Interessensgebiet der Person gehört (Krapp & Ryan, 2002). Dies führt möglicherweise zu einer Differenz zwischen Selbstwirksamkeitserwartung, tatsächlich ausgeführten Handlungen und realisierter Wirksamkeit. Dabei ist es sowohl möglich, dass das Zutrauen die eigenen Möglichkeiten übersteigt, als auch, dass Handlungen unterlassen werden, deren erfolgreiche Ausführung sich die Person nicht zutraut. Hier wird deutlich, an welcher Stelle die Selbstwirksamkeitserwartung bei Handlungen einer Person eine Rolle spielt: Nachdem eine Situation unter Berücksichtigung von personenbezogenen Kognitionen interpretiert wurde, allgemeine Wünsche wirksam werden können und Handlungsalternativen in die engere Auswahl gezogen werden, entscheiden Kontrollerwartungen und Kompetenzüberzeugungen wie die Selbstwirksamkeit, welche Ziele sinnvoll verfolgt und welche Handlungen tatsächlich auszuführen versucht werden.

Ähnlich wie personenbezogene Kognitionen nicht alleinerklärend fungieren, kann Verhalten in der hier vorausgesetzten Theorie nicht allein über den Grad der SWE erklärt werden. Das Konstrukt der SWE weist aber auch zusammen mit personenbezogenen Kognitionen einige blinde Flecken auf. Krapp und Ryan kritisieren am Konzept der SWE, das Konzept berücksichtige drei Faktoren nicht, die in anderen Modellen der Lernmotivationsmodellen einbezogen werden: a) eine qualitative Differenzierung verschiedener Formen der Lernmotivation, b) Ziele- und Inhaltsaspekte des Verhaltens und c) die Bedeutung emotionalen Erlebens im Prozessgeschehen (Krapp & Ryan, 2002). In der Tat folgen Zusammenhänge unter Berücksichtigung des Konzepts der SWE einfachen Prinzipien wie „eine hohe SWE führt bei gewähltem Ziel zu mehr Anstrengungsbereitschaft und Ausdauer“. Krapp und Ryan weisen darauf hin, dass die Qualität der Motivation (intrinsisch oder in einer der möglichen Formen extrinsisch) nicht berücksichtigt wird. Die Qualität der Motivation hat aber starken Einfluss auf das Verhaltensergebnis, die Qualität der erlebten Emotion und die Persistenz (Krapp & Ryan, 2002). Weiter bemerken Krapp und Ryan, dass in Banduras bahavioristischem Modell die Ziel- und Inhaltsaspekte des Handelns nicht berücksichtigt, sondern ausschließlich unter „Output“ die internalen und externalen Konsequenzen integriert sind (ebenda).

Aus Sicht der beiden Motivationspsychologen ist das Konzept der SWE schließlich zu kognitiv angelegt. Die so genannten „basic human needs“ (Kompetenzerfahrung, soziale Eingebundenheit, Autonomie) müssten vollständig bei der Erklärung menschlichen Verhaltens einbezogen werden. Es reicht nicht, sich nur auf die Kompetenzerfahrung zu beschränken, denn auch die beiden anderen seien unabdingbare Voraussetzung für das menschliche Wohlbefinden (Krapp & Ryan, 2002). Diese Kritik von Krapp und Ryan ist insgesamt nicht von der Hand zu weisen. Anders als in methodischen Studien zur SWE bei Lehrkräften wie von Schmitz und Kollegen wird die SWE in den hier besprochenen Handlungskompetenz-Modellen nicht allein als bedingende Variable der Handlungen herangezogen, sondern ist ein Teil vieler interagierender Variablen. Gleichzeitig kommt der SWE aber unbeachtete davon in der Regel die größte Aufklärungskraft zu.

Das Lehrer-Selbstkonzept und Unterricht

Nachdem die Konzepte der Selbstwirksamkeitserwartung und des Fähigkeitsselbstkonzepts in den vorherigen Abschnitten recht allgemein vorgestellt wurden, sollen in diesem Abschnitt konkrete Befunde zur SWE von Lehrkräften diskutiert werden. Im Zusammenhang mit Fragen nach der Professionalisierung von Lehrkräften und in der Lehrer-Experten-Forschung wird zwar auch das mathematisch-schulische Fähigkeitsselbstkonzept von Lehrkräften als relevante Variable angesehen und erhoben (Baumert et al., 2008), es werden allerdings keine Befunde berichtet. Die eingesetzten Fähigkeitsselbstkonzeptskalen dienen offensichtlich nur als Kontrollskala für die berichteten Variablen. Zur Selbstwirksamkeitserwartung vergleichbare Befunde sind folglich für das Fähigkeitsselbstkonzept von Lehrkräften nicht zu erwarten und können somit hier auch nicht berichtet werden.

Zusammenfassend kann die SWE als ein wichtiger Faktor in verschiedenen Kontexten von Bildungsinstitutionen und Bildungsprozessen angesehen werden. Insbesondere im Zusammenhang mit Fragen nach der Professionalisierung von Lehrkräften und in der Lehrer-Experten-Forschung wird die Selbstwirksamkeit von Lehrkräften als relevante Variable angesehen und erhoben. Die Selbstwirksamkeitserwartung spielt eine Rolle bei alltäglicher Unterrichtsgestaltung und bei speziellen Veränderungsprozessen. Lehrkräfte erleben eine geringere Belastung, wenn sie über eine hohe SWE verfügen (Schmitz, 2000). Eine hohe SWE hilft außerdem bei der Stressbewältigung (Jerusalem & Schwarzer, 1992) und der Möglichkeit, Reformprozesse ausdauernd und erfolgreich zu bewältigen (Jerusalem, 2002). Besonders die Überzeugung eigener Handlungsmächtigkeit begünstigt die Bereitschaft von Lehrkräften zu einer Veränderung, da sie das eigene Zutrauen stärkt, den Reformprozess erfolgreich abzuschließen. Wenngleich die Förderung von Selbstwirksamkeitsüberzeugungen nur einen kleinen Schritt zu erfolgreichen Reformen beitragen kann, gilt dies trotzdem als notwendig (Edelstein, 1998, 2002).

Höhere SWE führt darüber hinaus außerdem zu mehr Berufszufriedenheit und Verantwortungsbereitschaft bei Lehrkräften (Schunk 1995; Schunk/Meece 1992 – s. Edelstein 2002). Weniger selbstwirksame Lehrkräfte neigen laute Schwarzer und Jerusalem dazu, einen weniger komplexen Unterricht zu gestalten und kümmern sich seltener um Lernschwache, weil sie sich komplexere Unterrichtsplanung und ein differenziertes Eingehen auf Schülerinnen und Schüler nicht zutrauen. Der Unterricht von hoch selbstwirksamen Lehrkräften ist wesentlich herausfordernder und differenzierender, weil sie sich selbst mehr zutrauen und mehr Verantwortung übernehmen (Edelstein, 1998; Schwarzer & Jerusalem, 2002).

4.3 Lehrerpersönlichkeit als handlungsbestimmendes Element bei Rechenschaftslegung und Qualitätsentwicklung

Nach der Betrachtung von zentralen Lernstandserhebungen als Instrument der Neuen Steuerung in Kap. 2 werden nun diese in diesem Abschnitt im Zusammenhang mit der Lehrerpersönlichkeit auf einer individuellen Ebene dargestellt. Zentrale Lernstandserhebungen werden als Feedback-Intervention aufgefasst und dementsprechend werden Feedbackprozesse analysiert. Dies geschieht vor allem über die Feedback-Intervention-Theorie (FIT) von Kluger und DeNisi (1996) bzw. die Feedback-Theorie von Hattie und Timperley. Daraus lassen sich einzelne Variablen der Persönlichkeit der gefeedbackten Person, der Situation und des Feedbacks selbst ableiten, deren Bedeutung und dazu getätigte Befunde vorgestellt werden. Anschließend werden dieser Sammlung einzelner Variablen die Idee von Typenmodellen zu Innovationen und im Sinne des Evaluationszykluses nach Helmke (vgl. 2.2.4) gegenüber gestellt. Dieser Ansatz der Typenmodelle wird in dieser Arbeite auch in Studie B aufgegriffen. Abschließend folgt in diesem Abschnitt ein kurzer Überblick über Befunde zur deutschen Rezeptionsforschung. Zu Beginn hingegen soll die Wirkung von Feedback auf Individualebene erklärt werden. Ein Augenmerk liegt dabei auf der Reaktanz gegenüber von Feedbackprozessen.

4.3.1 Zentrale Lernstandserhebungen als Feedback-Intervention

Die bereitgestellten Ergebnisse aus zentralen Lernstandserhebungen können als Feedback-Intervention betrachtet werden und es ist prinzipiell naheliegend, dieser Feedback-Intervention eine Wirkung auf die Unterrichtsqualität zu unterstellen (Ditton & Arnoldt, 2004a). Voraussetzung für Veränderungsprozesse ist ein wahrgenommenes Defizit. Dieses kann durch eine Ergebnisrückmeldung veranschaulicht werden. Administrativ gewünschte

Schul- und Unterrichtsentwicklungsprozesse scheitern bisher häufig an fehlendem Problembewusstsein der Lehrkräfte und Schulleitungen (Visscher, 2008), aber ebenso an Schwierigkeiten, die angebotenen Daten zu verstehen (vgl. unten). Feedback kann – unabhängig der zugrunde liegenden Theorie - als einer der einflussreichsten Faktoren für das Lernen und hohe Leistungsergebnisse angesehen werden (Hattie & Timperley, 2007). Es stellt somit eine wichtige Unterstützung für die Innovationskompetenz bzw. Innovationsaufgabe der KMK-Standards zur Lehrerbildung dar. Ohne Feedback werden Innovationsproesse selten angestoßen oder verlaufen unreflektiert.

Ein deutlich negativer Aspekt von Veränderungsprozessen stellt Widerstand gegen Evaluationen im schulischen Kontext dar. Widerstand gegen Evaluationen kann sowohl individueller Natur als auch die überpersonelle Haltung einer Fachgruppe oder ganzer Kollegien sein und kann u.a. aus der Angst vor Überlastung oder aus einer generellen Skepsis gegenüber Veränderungen resultieren. Es bieten sich drei psychologische Theorien als Erklärung an: (1) Die Reaktanztheorie nimmt an, dass Widerstände gegenüber Veränderungen aus der Angst resultiert, Freiheit und Handlungsalternativen und damit Teile der Professionalität zu verlieren. Die Größe des Widerstands hängt danach von der Wichtigkeit der bedrohten Freiheit, des Umfangs (Beisp.: Einführung von neuen Steuerungsinstrumenten und neuem Lehrplan innerhalb kurzer Zeit) und Stärke der Veränderung (zwischen Empfehlung und unumstößliche Forderung). (2) Nach der Theorie der kognizierten Kontrolle (TdkK) entsteht Widerstand, wenn Personen Zustände und Ereignisse nicht mehr kontrollieren, erklären und vorhersagen können. Besonders sich ihrer Kompetenz unsichere Personen sind betroffen. Widerstand lässt sich nach der TdkK verhindern, wenn über die Veränderungen ausreichend informiert und die Betroffenen in den Prozess einbezogen werden. (3) Die Selbstaufmerksamkeitstheorie knüpft an ein Problem an, das auch in der FIT von Kluger und DeNisi thematisiert wird. Personen kann es unangenehm sein, auf Schwächen hingewiesen zu werden (Steins, 2009).

Kurz gesagt können sowohl die befürchteten und durch zentrale Lernstandserhebungen angeregten Veränderungen in Unterricht und Schule als auch die Einführung der Lernstandserhebungen selbst zu Widerstand unter Lehrkräften führen. Mögliche Folgen können beispielsweise offene Sabotage (durch „Mogeln“ bei zentralen Lernstandserhebungen) oder „Dienst nach Vorschrift“ (Ergebnisrückmeldungen werden nur scheinbar rezipiert und reflektiert) sein. Als latenter Widerstand kann Testcoaching bezeichnet werden, wenn damit bewusst die Aussagekraft der Ergebnisse von Leistungsmessungen gefährdet wird.

Die zweite Sichtweise auf Testcoaching als Vorbereitung auf Lernstand8 nimmt zentrale Lernstandserhebungen folglich als systematisches Feedbackangebot mit möglichen positiven wie negativen Wirkungen in den Blick. Im dritten Kapitel ist erläutert worden, inwiefern Testcoaching nicht nur eine konkrete Form der Unterrichtsqualität darstellt, sondern auch welcher Unterrichtsentwicklung es bedarf, um dies in den Unterricht zu integrieren. In welchem Ausmaß die beabsichtigte Wirkung gelingt, hängt von der Qualität des Feedbacks

ab und ist Gegenstand der sich ausweitenden Rezeptionsforschung (s. 4.3.5). Auch bisher vorgelegte Typisierungsansätze (s. 4.3.4) zeigen Anknüpfungspunkte zum Zugang der Rezeptionsforschung.

Die allgemeine Bedeutung von Feedback lässt sich mit behavioristischen, kognitivistischen und konstruktivistischen Theorien erklären. Die kognitivistische Modelle wie die Zielsetzungstheorie (nach Locke & Latham, 1990), die Kontrolltheorie (nach Carver& Schleicher, 1981, 1982) oder die Handlungstheorie (Frese & Zapf, 1994), aber auch konstruktivistische Lehr-Lern-Theorien setzen voraus, dass mit den angebotenen Feedback-Informationen bewusst umgegangen wird.[83] Neben der Verständlichkeit der Feedback-Information und den individuellen Verarbeitungsmöglichkeiten des Empfängers ist somit als dritte Komponente seine Motivation, Feedback zu empfangen und zu verarbeiten, zu berücksichtigen.

Zu der Wirkungsweise von Feedback existiert ein weites Feld an Studien, die jeweils einzelne Aspekte dieses Komplexes untersuchen. Typische Befunde sind dazu im Abschnitt 4.3.3 skizziert. Die den Studien zugrunde liegenden Theorien decken aber immer nur einen kleinen Teil der bisherigen Erkenntnisse ab. Zum Teil sind diese Befunde sogar nicht miteinander vereinbar (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Narciss, 2006). Eine komplette Feedback-Theorie ist die kognitivistisch Feedback-Interventions-Theorie (FIT) von Kluger und DeNisi aus dem Jahre 1996. In der FIT werden alle bis dahin bekannten Einflussvariablen und beobachteten Prozesse klassifiziert und die Wirkungsmöglichkeit von Feedback zur Leistungsförderung wird über die Hierarchie von Aufmerksamkeit erklärt. Dieser umfassende Anspruch macht die FIT auch im Zuge der Forschung über den Umgang mit den Ergebnissen aus zentralen Lernstandserhebungen reizvoll. Sie wird daher im zweiten Abschnitt dieses Unterkapitels dargestellt. Dabei beschränkt sich die Darstellung allerdings auf die abstrakten Aspekte der FIT, einzelne Ergebnisse der ihr als Basis dienenden Metaanalyse werden ausgespart. Hattie und Timperley haben 2007 eine konstruktivistisch Feedback-Theorie vorgestellt, die in Teilen auf Kluger und DeNisi aufbaut, aber vorwiegend auf das Lernen ausgerichtet ist, während die FIT von Kluger und DeNisi Feedback analog zu kybernetischen Prozessen beschreibt und die folgenden Prozesse analysiert. Auch ihr Modell wird in Abschnitt 4.3.2 vorgestellt, um der zweiten Sichtweise (zentrale Lernstandserhebungen als Teil eines Lernprozesses für Lehrkräfte) gerecht zu werden. Der Grund dafür, dass die Rückmeldungen auf die Lernstandserhebungen zwar als Feedback, aber nicht zwingend als Teil eines Lernprozesses durch gegebenes Feedback aufgefasst werden, liegt in den Ergebnissen einer von Vollmeyer und Rheinberg durchgeführten Studie zur Wirkung von angekündigtem Feedback (2005). Ein Ergebnis ihres Versuchs ist eine vorwegnehmende Wirkung von Feedback. Die Versuchsteilnehmer nutzten in dem Versuch von Beginn an systematischere Strategien, um das ihnen gestellte Rätsel zu lösen, wenn ihnen angekündigt wurde, nach einigen Versuchen ein Feedback über ihre Leistung zu erhalten (Vollmeyer & Rheinberg, 2005). Diese vorwegnehmende Wirkung von Feedback wird in grundlegender

[83] Dies kann allerdings auch ein bewusstes Ignorieren meinen. - Für eine ausführlichere Klassifikation der verschiedenen Feedbackansätze s. Narciss (2006).

Form auch beim Teaching to the Test unterstellt und daher auch in dieser Arbeit angenommen.

Feedback als Begriff

Feedback ist allgemein eine bestimmte Art von Information über eine mögliche Diskrepanz zwischen einem Ist-Zustand und Soll-Zustand und setzt somit ein vorher definiertes Ziel voraus (Frese & Zapf, 1994). Der Zweck der Feedback-Information ist die Regulation eines Prozesses (Narciss, 2006), in der Psychologie sind damit Verhalten und Lernprozesse gemeint. Narciss unterscheidet informatives Feedback, welches sich auf die Lösung einer Aufgabe bezieht und auf die korrekte Lösung abzielt, von motivierendem Feedback, das Sanktionen beinhalten kann und häufig die individuelle Bezugsnorm zugrunde legt, und von internem Feedback, also vom Lernenden selbst wahrgenommene Informationen (ebenda). Lernstandserhebungen als Instrument der Rechenschaftslegung und Qualitätsentwicklung trennen den Feedbackprozess in den Ist-Soll-Vergleich als Rechenschaftslegung bzw. Qualitätsüberprüfung und die resultierende Handlung als Qualitätsentwicklung.

Das Feedback kann eine Übereinstimmung von Ist- und Sollzustand oder eine Diskrepanz zwischen beiden aufweisen. Von negativem Feedback wird im eigentlichen Sinne gesprochen, wenn der Ist-Soll-Vergleich eine Abweichung ergibt, eine Übereinstimmung wird als positives Feedback bezeichnet. Diese Sichtweise wird von den kognitivistischen Feedbackansätzen geteilt (Müller, 2008). In Lernprozessen und bei bestimmten Verhaltensprozessen wird aber auch ein Übererfüllen des Ziels mit positivem Feedback berücksichtig.

Prinzipiell sind verschiedene Handlungen möglich, um mit diesem etwaigen Missverhältnis umzugehen: (i) Das Verhalten kann so angepasst werden, dass die Diskrepanz gemindert wird. Dies entspricht in der Regel der Intention von Feedback-Interventionen. Dazu sind aber ausreichend Motivation für die Aufgabe und klare Zielvorstellungen notwendig (Maier, 2009). (ii) Die Standards können angepasst werden. Dies ist die intendierte Maßnahme, wenn nicht das Ziel, sondern die gefeedbackte Person im Vordergrund steht, und falsche Ansprüche verdeutlicht werden sollen. (iii) Die Standards können auch komplett verworfen werden. Dies erklärt Maier z.B. mit fehlender Hoffnung auf Erfolg (2009). (iv) Die Wahrheit des Feedbacks kann angezweifelt werden. Dies kommt bei negativem Feedback häufiger vor als bei positivem Feedback, fanden Kluger und DeNisi. Der Befund gilt aber nicht kulturübergreifend, sondern nur für die westlichen, weniger kollektiven Gesellschaften (Kluger & DeNisi, 1996). Es ist zusätzlich nicht in Einklang zu bringen mit dem Selbstkonsistenzmotiv (s.u.) (v) Die gefeedbackte Person kann die Situation schließlich auflösen, indem sie sich der Feedbacksituation entzieht. Die letzten drei Möglichkeiten können nicht als mögliche Intention von Feedback-Interventionen betrachtet werden, trotzdem stellen sie typische Handlungsmuster im Umgang mit Feedback dar. Kombinationen aus den fünf Strategien sind ebenfalls möglich. Die grundlegenden Prinzipien

dieser fünf Handlungsalternativen werden in Theorien zur Dissonanzreduzierung von Fremd- und Selbstbild (Aronson, Wilson & Akert, 2004; Festinger, 1957; Fischer, Frey, Peus & Kastenmueller, 2008) und mit der Attributionstheorie erklärt.

Attributionsstile

Grundsätzlich können Leistungen internal und external attribuiert werden. Außerdem können Leistungen auf zeitlich stabile und zeitlich variable Faktoren zurückgeführt und als kontrollierbar oder unkontrollierbar angesehen werden. Bromme und Rheinberg berichten von verschiedenen Studien, nach denen Lehrkräfte häufiger Schülerleistungen external attribuieren, also den Fähigkeiten, Anstrengung oder Begabungen der Schülerinnen und Schüler zuschreiben (Bromme & Rheinberg, 2006). Tresch berichtet über vorwiegend einen typisch optimistischen Attributionsstil bei Lehrkräften. Sie attribuierten die Ergebnisse ihrer Schüler und Schülerinnen bei der Schweizer Studie Check 5 insgesamt eher internal, wenn diese überdurchschnittlich gut waren, und eher external bei unterdurchschnittlichen Resultaten (Tresch, 2007). Dies wird allgemein als *Self-serving Bias* bezeichnet (Musch & Bröder, 1999). Beide Befunde konnte auch Schneewind replizieren. In einer von ihr durchgeführten Interviewstudie zum Projekt BeLesen zeigte sich, dass Lehrkräfte die rückgemeldeten Testergebnisse auf den individuellen Schülerinnen und Schüler bezogen. Eine Wahrnehmung auf Klassenebene fand nicht statt (Schneewind, 2007). Auch die Lehrer-Befragung zu VERA3 in den Jahren 2004 bis 2008 zeigt ähnliche Ergebnisse. Koch konnte aber in einer quasi-experimentellen Interventionsstudie zeigen, dass ein möglicher Grund in unzureichenden statistischen Fähigkeiten der Lehrkräfte liegen könnte, die Ergebnisrückmeldung korrekt zu interpretieren. Wenn Lehrkräfte eine Statistikschulung absolvierten, hat sich bei der Ergebnisrückmeldung ihr Fokus von den Individualergebnissen zu Gruppenanalysen verschoben (Koch, 2011). Kohler fand in ihrer Rezeptionsstudie zum Umgang mit den Ergebnissen der deutschen TIMSS-Ergebnisse einen Zusammenhang zwischen externaler Attribution und einer eher geringen zudachten Bedeutung der Ergebnisse durch Lehrkräfte (Kohler, 2009). Als externale Ursachen können auf überindividueller Ebene außerdem soziale Bedingungen (soziale Herkunft, ethnische Zusammensetzung der Klasse) angeführt werden. Handelt es sich um eine Leistungsmessung mit externen Aufgaben, werden auch diese als Ursache von Lehrkräften für die Erklärung erwartungswidriger Testleistungen herangezogen.

Kühle konnte dabei in einer Befragung der Fachgruppenleiter zu Lernstand8 zeigen, dass die Ursachenzuschreibungen schulformabhängig sind. Während Lehrkräfte an Gesamt- und Hauptschulen erwartungswidrige Ergebnisse häufiger vor allem der Schülervariable zuschrieben, griffen Gymnasiallehrkräfte häufiger als die Gesamt- und Hauptschullehrkräfte auf die Art der Aufgabenstellungen als Erklärung zurück (Kühle, 2010). Die sozialen Bedingungen werden vergleichsweise selten für die Erklärung schlechter Leistungen herangezogen, an Gesamt- und Hauptschulen allerdings etwas häufiger als an Gymnasien und Realschulen. Nur Gesamtschullehrkräfte gaben in Kühles Befragung häufig interne

Ursachen (Unterrichtsgestaltung, methodisch-didaktisches Vorgehen, verwendete Unterrichtsmaterialien, Lehrerkompetenz) an, die Lehrkräfte anderer Schulformen sagen hingegen in den internen Ursachen die geringste Erklärungskraft. Kühle kommt insgesamt zum Schluss, dass der Attributionsstil die Reflexions- bzw. Veränderungsbereitschaft nicht stark beeinflusst, ein externer Attributionsstil aber zu einer geringeren Reflexions- bzw. Veränderungsbereitschaft führt (Kühle, 2010). An dieser Stelle muss allerdings noch einmal darauf hingewiesen werden, dass nur Fachgruppenleiter befragt wurden. In der Regel können diese über das tatsächliche Unterrichtsgeschehen einzelner Kollegen keine Aussage treffen.

Musch und Bröder stellen die Erklärung des *Self-serving Bias* in Frage. Sie kommen aufgrund eigener Studien zu dem Schluss, dass das Attributionsmuster vom Fähigkeitserleben abhängt. Menschen mit häufigen Erfolgserlebnissen schreiben Misserfolge Ausnahmebedingungen zu, die eher external zu verorten sind, Menschen mit vielen Misserfolgen handeln entsprechend bei Erfolgen (Dauenheimer, Stahlberg, Frey & Petersen, 2002; Musch & Bröder, 1999). Beides scheint erst einmal unter dem Aspekt der nach Hattie und Timperley geringen Zahl an systematischen Feedbackhinweisen an Lehrkräfte (Hattie & Timperley, 2007) und den dadurch wenigen Möglichkeiten für Lehrkräfte, ein elaboriertes Fähigkeitsselbstbild zu erlangen, widersprüchlich und kann ohne weitere Forschung an dieser Stelle nicht aufgelöst werden. Eine mögliche Erklärung könnte sein, dass der *Self-serving Bias* in den Situationen durchaus zum Tragen kommt, in denen kein (korrigierendes externes) Feedback vorliegt.

Grundsätzlich muss nach der Attributionstheorie angenommen werden, dass negatives Feedback nur dann zu einer Veränderungsreaktion führt, wenn die Gründe internal und instabil attribuiert werden und als kontrollierbar betrachtet werden (Ilgen & Davis, 2000). Mit Musch und Bröder müssten die Lehrkräfte zusätzlich außerdem selbst ein positives Selbstbild besitzen.

4.3.2 Feedback-Theorien

Die Feedback-Interventions-Theorie nach Kluger und DeNisi

Kluger und DeNisi definieren eine Feedback-Intervention als *eine Handlung einer externen Person oder mehreren externen Personen, durch die einer anderen Person Wissen über einen ihrer Tätigkeitsaspekte eingeräumt wird* (1996). Die Definition verbindet folglich drei Begriffe miteinander: (1) die externe Person, die das Feedback anbietet, (2) die Person, der das Feedback angeboten wird und (3) eine von der gefeedbackten Person durchgeführte Tätigkeit. Die Theorie befasst somit ausdrücklich nicht mit Feedback zu einer Person (Kluger & DeNisi, 1996).

In einer von ihnen durchgeführten Meta-Analyse mittels 131 Studien zu Wirkungen von Feedback-Interventionen auf die Leistung kommen sie zu dem Ergebnis, dass von 607 berichteten Effekten ein Drittel eine negative Effektstärke und ungefähr die Hälfte eine Effektstärke nahe bei null aufwiesen und somit nur ein kleiner Teil der Studien bemerkenswerte positive Effekte vorweisen konnte. Obwohl sich in den Studien, die Kluger und DeNisi für ihre Metaanalyse heranzogen, für zumindest die ersten vier Verarbeitungsstrategien durchaus empirische Belege finden lassen, kritisieren sie die Unterkomplexität der Erklärungen. Die Studien können ihrer Ansicht nach erstens nicht erklären, wie sich die Multidimensionalität eines jeden Standards auswirkt. Standards ergeben sich aus den Bezugsnormen, mit denen Leistung gemessen werden kann. Neben den drei bekannten (s.o.) Bezugsnormen ist eine Bewertung demnach darin möglich, ob eine auf sich selbst bezogene Leistung oder eine Leistung als Mittel zu etwas erbracht werden soll. Möglich ist nicht nur eine abweichende Ansicht über die eine als Maßstab zu verwendenden Bezugsnorm, sondern auch eine Kombination aus mehreren. Dadurch löst ein Feedback teilweise verschiedene Emotionen gleichzeitig aus, die summiert werden. Vor allem kann es zu einem Kommunikationsproblem kommen, wenn die Ziele der gefeedbackten Person nicht ausreichend bekannt sind oder sich widersprechende Rückmeldungen ergeben. Zweitens weisen nach Kluger und DeNisi die Studien Erklärungslücken auf, wenn es darum geht zu begründen, wieso Feedback bei komplexen Tätigkeiten eher geringe oder gar negative Erfolge aufweist und warum Zufriedenheit und Erregung, zwei der Leistung förderliche Gefühlszustände, einerseits durch die Richtung des Feedbacks (Zufriedenheit – je höher je positiver) und andererseits durch die Diskrepanz-Größe (Erregung – je höher je größer die Differenz) bedingt werden.

Kluger und DeNisi sehen daher die Notwendigkeit eine spezielle Feedback-Interventions-Theorie zu postulieren, die auf fünf Annahmen beruht: (1) Verhaltensregulation findet mittels Abgleich von Feedback zu Zielen und Standards statt, welches sich sowohl mit der Kontrolltheorie (Ziel hier: Minderung der Ist-Soll-Diskrepanz) und der Zieltheorie (Ziel hier: Erreichen des gesetzten Ziels) erklären lässt. (2) Diese Ziele und Standards sind hierarchisch angeordnet. (3) Die Aufmerksamkeit eines Menschen ist beschränkt, sodass nur bestimmte Ist-Soll-Differenzen ausreichend Aufmerksamkeit erhalten, um eine Verhaltensänderung anstoßen zu können. (4) Ohne Feedback-Intervention ist die Aufmerksamkeit auf eine mittlere Hierarchieebenen (assoziiert mit Task-Motivation Processes) zwischen automatisch ablaufenden Handlungen (assoziiert mit dem Task-Learning Processes) und der Ebene des Selbst (assoziiert mit dem Meta-Task Processes) gerichtet. (5) Feedback-Interventionen lenken die Aufmerksamkeit auf andere Stellen und regen dadurch zu Verhaltensmodifikationen an (Kluger & DeNisi, 1996).

Die FIT von Kluger und DeNisi betrachtet diese fünf Auswege unter den drei Prozessen Task-Learning Process, Task-Motivation Process und Meta-Task Process. Feedbackschleifen der ersten Hierarchieebene, also auf Tätigkeitsdetails beschränkte Feedbackschleifen, sind Feedbackschleifen der beiden anderen, Feedbackschleifen der Tätigkeitsebene nur den Feedbackschleifen der Selbstebene untergeordnet. Wie in den fünf Grundannahmen

aufgelistet ist der Task-Motivation Process stets die Ausgangslage. Wird bei einem Feedbackvorgang Ist- und Soll-Zustand verglichen und ein Defizit erkannt, führt dies in der Basistheorie erst einmal zu erhöhter Anstrengung (bei positivem Feedback folgt keine Veränderung oder die Ziele werden höher angesetzt). Von dort kann die Aufmerksamkeit auf den Lernprozess oder das Selbst gelenkt werden, wenn das Feedback negativ ausfällt und gleichzeitig eine höhere Anstrengung kein besseres Feedback einbringt bzw. erwarten lässt.

Wird nun weiterhin das Erreichen des Ziels als relevant angesehen und glaubt die Personen, die Differenz durch eine Verhaltensänderung beheben zu können, wechselt die Person in den Lernprozess. Diese Aufmerksamkeitsänderung kann auch dadurch begünstigt werden, dass das Feedback sich auf bestimmte Aspekte richtet. Der Lernprozess ist schließlich erfolgreich, wenn selbst ein Ausweg gefunden wird, um die Ist-Soll-Differenz auszugleichen oder das Feedback entsprechende Hinweise beinhaltet. Anderenfalls bleibt ein Lerneffekt aus, kann es zu negativen Lerneffekten kommen (wenn vermeintliche Auswege entdeckt wurden) oder die Feedbacksituation wird gemieden (wenn das Ziel nicht wichtig genug erscheint).

Erweist sich nach einem negativen Feedback eine intensivere Anstrengung als wirkungslos und ist auch keine Verhaltensänderung Option zur Defizitbeseitigung in Sicht, wechselt die Person in den Meta-Task Process. Dieser Aufmerksamkeitswechsel kann unter vier verschiedene und voneinander unabhängige Erklärungsansätzen betrachtet werden, nämlich die Bemühungen, die Dissonanz zwischen Fremd- und Selbstbild zu reduzieren, und affektive Prozesse bzw. eine erhöhte Selbstaufmerksamkeit und der Abzug von kognitiven Ressourcen für die eigentliche Tätigkeit (Kluger & DeNisi, 1996).

Unter dem Blickwinkel der erhöhten Selbstaufmerksamkeit betrachtet ist der Meta-Task Process durch einen entsprechenden expliziten Hinweis im Feedback ausgelöst und möglicherweise bewusst angestrebt gesehen worden. Dies ist beispielsweise der Fall, wenn sich das Feedback um die Einhaltung einer sozialen Norm dreht oder auf Ziele höherer Ordnung bezieht („Du machst einen guten Eindruck.“). Sind hingegen Leistungssteigerungen oder die Änderung eines Verhaltens beim Ausführen einer Tätigkeit das Ziel, so bedeutet eine erhöhte Selbstaufmerksamkeit gleichzeitig immer auch weniger Aufmerksamkeit für die Aufgabe selbst. Auch dies kann gewollt sein, wenn die Tätigkeit keine große kognitive Leistung erfordert. Verstärkendes Feedback, das auf das Selbst gerichtet ist, kann zu einer Motivationssteigerung führen. Ist die Tätigkeit hingegen komplex, wie beispielsweise die Unterrichtstätigkeit, werden der Tätigkeitsausführung wichtige Ressourcen entzogen.

Der Umgang mit Feedbackhinweisen auf der Selbstebene kann als affektiver Prozess oder aber als Reduzierung der kognitiven Dissonanz geschehen. Als affektiver Prozess erklärt er, wie Ängste, Unwohlsein mit erhöhter Erregung und auch Freude und Stolz durch bewertende Hinweise im Feedback verursacht werden und wie daraus höhere Risikobereitschaft oder eine Aversion gegen derartige Tätigkeiten folgen können. Die Reduzierung der kognitiven Dissonanz hingegen betrachtet diese Prozesse als kognitive, mit anderen Worten durch Nachdenken vorgenommene Prozesse, die zu bewussten

Entscheidungen über nachfolgende Handlungen führen. Diese Arbeit folgt diesem Modell der bewusst vorgenommenen Entscheidungen.

Die Feedback-Theorie von Hattie und Timperley

Hattie und Timperley definieren Feedback als *Information eines Akteurs (beispielsweise eine Lehrkraft, die Erziehungsberechtigten, ein Lösungsbuch oder man selbst) über die Leistung einer Person oder ihres Verständnisses.* Dabei ist die Feedback erhaltende Person erst einmal nur passiv existent, Hattie und Timperley nehmen aber Bezug auf Winne und Butler (1994), die wiederum Feedback als Information definiert haben, die man annehmen, sich merken kann und die neu Erkenntnisse liefern kann. Die vorher passive Person, die in der Definition von Kluger und DeNisi zwar nicht passiv, aber durch die Art des Feedbacks, die Tätigkeit und die Situation bzw. ihre Persönlichkeit zu einer Handlung veranlasst wurde, wird dadurch zu einer Person mit freiem Willen. Das angebotene Feedback kann folglich akzeptiert werde, es kann aber genauso gut modifiziert oder sogar zurückgewiesen werden (Hattie & Timperley, 2007).

Gutes Feedback beantwortet nach Hattie und Timperley in jedem Fall drei Fragen: (1) Wie lautet das Ziel? (2) Wie gut habe ich mich bisher auf dem Weg zum Ziel geschlagen? (3) Was kann ich tun, um dem Ziel näher zu sein als bisher? Die dritte Frage wandelt das Feedback in ein Feed-forward, weil hierin Informationen enthalten sind, die sich auf noch zu tätigende Schritte des Tätigkeitsprozesses beziehen.

Nach dem Feedback-Model von Hattier und Timperley wirkt Feedback immer gleichzeitig auf bis zu vier Ebenen: (a) Die Aufgabenebene (FT für Feedback to Task) gibt an, wie gut eine Aufgabe verstanden wurde bzw. welche Leistung erbracht wurde. (b) Die Prozessebene (FP für Feedback to Process) umfasst den Hauptprozess, mit dem die Aufgabe bearbeitet wird bzw. die Leistung erbracht wird. (c) Die Selbstregulationsebene (FR für Feedback to Regulation) kombiniert sowohl Selbstevaluationsstrategien als auch den Motivationsbereich, der die gefeedbackte Person anregt weiter zu arbeiten. (d) Mit der Selbstebene (FS für Feedback to Self) ist der Teil des allgemeinen Bereichs des Selbstkonzepts angesprochen. Die Effektivität des Feedbacks im Sinne einer Leistungssteigerung oder eines Lernprozesses ist für die Aufgabenebene am größten und für die Selbstebene am niedrigsten

Im Vergleich zur FIT von Kluger und DeNisi kommt damit eine vierte Ebene ins Spiel. Die FT-, FP- und die FS-Ebene lassen sich sehr gut mit den drei Ebenen der FIT assoziieren, wenngleich dort die Unterscheidung zwischen der mittleren und der unteren Aufmerksamkeitsebene über den Grad der Spezifikation des Feedbacks vorgenommen wird und nicht zwischen Tätigkeit und Prozess unterschieden wird. Die vierte Ebene des Feedback-Modells nach Hattie und Timperley stellt die Selbstregulation heraus und ist damit typisch für Theorien aus der Gruppe der Lehr-Lern-Theorien (Maier, 2009).

Nach Hattie und Timperley sind auf dieser Ebene fünf Variablen zu verorten, die bei Feedback als Mediatorvariablen fungieren: (1) Capability to Self-assessment kombiniert die Fähigkeit, sein eigenes Können und Wissen richtig einzuschätzen, mit der Fähigkeit planvoll aus Fehlern zu lernen und das eigene Vorgehen zu analysieren. (2) Der Willingness to invest Effort into Seeking and Dealing with Feedback Information meint die Bereitschaft, den notwendigen Einsatz für Feedbacksituationen zu aufzubringen. Es entstehen physische Kosten für das Erreichen des Feedbacks, emotionale Kosten durch die Beurteilung des eigenen Tuns durch andere und kognitive Kosten aus der Verarbeitung der Feedbackinformationen. (3) Mit Degree of Confidence ist das Ausmaß der Selbstzufriedenheit und die Erfolgserwartung gemeint. Umso höher diese ausfällt, umso größer muss die zurückgemeldete Abweichung sein, damit der Feedbackinformation ausreichend Aufmerksamkeit geschenkt wird. (4) Wichtig ist außerdem der Attributionsstil der gefeedbackten Person. Je mehr eine Person die eigene Leistung sich selbst zuschreibt, umso empfänglicher ist sie für Feedbackinformationen zu diesen Leistungen. (5) Schließlich ist für die Weiterentwicklung das Level of Proficiency at Seeking-help relevant, das Ausmaß der Fähigkeit also, wie stark eine Person in der Lage ist, sich Unterstützung und Hinweise zur Weiterentwicklung oder Verbesserung zu suchen.

4.3.3 Eine Betrachtung einzelner relevanter Variablen für den Feedbackprozess

Wenn es um die Wirkung von Feedback geht, müssen zwei Fragen unterschieden werden: (1) Wie muss Feedback gestaltete sein, damit es zu einem beabsichtigten Lernprozess beiträgt? und (2) Wie wirken sich die bereitgestellte Feedbackinformation und die Ankündigung von Feedbackinformationen aus? Diese Differenzierung ist nötig, da Feedback einerseits unwirksam sein kann, andererseits aber auch unbeabsichtigte Wirkungen erfolgen können und dies negative Konsequenzen nach sich ziehen kann. Studien zur Wirkung von Feedback thematisieren häufig nur die erste Frage. Für eine detailliertere Analyse können hier analog zum Evaluationskreislauf von Helmke (vgl. 2.2.4) die Rezeption der Feedbackinformation, die Reflexion und die abgeleitete Handlung (Aktion) unterschieden werden (Helmke, 2004). Die meisten Studien verzichten allerdings auf diese Differenzierung und vergleichen lediglich den Lernerfolg, folglich eine Form der Aktion. Befunde zur zweiten Frage sind äußerst rar, da unbeabsichtigte Effekte nur seltenvorher abgeschätzt werden können und daher auch keine Messinstrumente eingesetzt werden, um diese erfassen zu können.

Die Studie von Vollmeyer und Rheinberg ist ein Beispiel für eine unbeabsichtigte Wirkung einer angekündigten Feedbackinformation, die allerdings positive Konsequenzen hatte und die bereits vorab zur erwarten waren, wenn auch zu einem späteren Zeitpunkt im Experiment. Die Teilnehmer am Experiment wurde komplexe Simulationen am Computer vorgelegt. Durch strategisches Vorgehen sollten sie die Zusammenhänge des simulierten

biologischen Organismuses erforschen und angeben. Der Experimentalgruppe wurde zu Beginn nach jeder Runde ein Feedback zu ihrem Vorgehen versprochen. Es zeigte sich, dass die Experimentalgruppe durchschnittlich mit systematischeren Strategien agierte als die Kontrollgruppe. Dies geschah aber anders als erwartet nicht erst nach dem ersten oder einem späteren Feedback, sondern schon von Beginn an (Vollmeyer & Rheinberg, 2005).

Lehrkräfte erhalten durch das Verhalten ihrer Schülerinnen und Schüler und durch die von Schülerinnen und Schülern erarbeiteten Lösungen zu von den Lehrkräften gestellten Aufgaben zwar stetig Feedbackinformationen. Dabei handelt es sich allerdings um internes Feedback, welches vorwiegend unsystematisch gewonnen wird. Systematisches externes Feedback könnte durch Kollegen, Schülerbefragungen oder auch durch die Schulleitung erfolgen, kommt aber nur sehr selten vor. Fünfzehn Prozent der Lehrkräfte geben an, selten eine Rückmeldung von der Schulleitung zu erhalten, jede vierte Lehrkraft empfindet die Rückmeldungen durch die Schulleitung als nicht ausreichend (Dick, 2006). Auch Schüler und Schülerinnen erhalten andersherum nur selten auf den Lernprozess ausgerichtetes Feedback. Stattdessen stehen summative Rückmeldungen im Vordergrund. Hattie und Timperley sehen in Schulleistungsmessung eine Möglichkeit, mehr Anlässe für Feedbacksituationen zu schaffen, in denen FT, FP und FR an Schüler und Schülerinnen gegeben werden (Hattie & Timperley, 2007). Noch weniger erhalten Lehrkräfte selbst ein derartiges Feedback, sodass LSE hier eine der wenigen Möglichkeiten bietet, ein erst einmal nicht auf das Selbst der Lehrkraft abzielendes Feedback zu generieren.[84] Auch für die Schulleitung könnten die Lernstandserhebungen Anlass für eine systematische Rückmeldung sein.

Nach Kluger und DeNisi können bei der Wirkung von Feedbackinterventionen drei Variablenklassen unterschieden werden: (i) die Facetten der Feedbackbotschaft, (ii) die ausgeführte Tätigkeit, (iii) die Situations- bzw. Persönlichkeitsvariablen (Ditton & Arnoldt, 2004).

Die Facetten der Feedbackbotschaft sind Hinweisreize, die die Aufmerksamkeit auf eine der drei Hierarchieebenen lenken und die damit assoziierten Prozesse begünstigen können. Forschung zu diesen Hinweisreizen umfassen bezogen auf Rückmeldungen aus zentralen Lernstandserhebungen u.a. die Rückmeldearten wie sie in den Arbeiten von Schneewind (2007), Maier (2009) und Müller (2010) untersucht wurden.[85] Grundsätzlich sind Hinweisreize, die auf die Tätigkeitsebene oder die Detailebene abzielen günstig (Jakobs, 2008). Auf das Selbstbild zielende Aufmerksamkeitsanreger sind in den Rückmeldungen zu zentralen Lernstandserhebungen zweifach vorhanden: Die als Vergleichsmaßstab zurückgemeldeten sozialen Vergleichswerte mit anderen Klassen der Schule und Schulen des gleichen Standorttyps können die Aufmerksamkeit auf das Selbst richten, aber Diskussion der Ergebnisse mit Fachkollegen und in Fachgruppen können problematisch sein. Müller

[84] Zur Frage, ob die rückgemeldeten Ergebnisse von Lernstandserhebungen summativ oder formativ sind vgl. Abschnitt 2.2

[85] Zu einzelnen Befunden s. Abschnitt 4.3.5 zur Rezeption von Ergebnissen aus SSL im deutschsprachigen Raum.

kritisiert hieran, dass die Rückmeldungen zu den zentralen Lernstandserhebungen indirekt und zu abstrakt sind, ohne inhaltlich in die Tiefe zu gehen (Müller, 2010).

Wichtig ist in jedem Fall für eine positive Wirkung auf den Lernprozess, nicht nur bewertende Informationen, sondern auch Hilfestellungen zu geben. In der Meta-Analyse von Kluger und DeNisi wiesen rein wertende Rückmeldungen (negativ und positiv) teilweise negative Effektstärken auf (Kluger & DeNisi, 1996; Maier, 2009). Dabei sollten die Informationen über korrekte Tätigkeitsteile gegenüber denen zu falschen im Vordergrundstehen (Hattie & Timperley, 2007).

Laut Hattie und Timperley kann zur Zeitspanne zwischen Tätigkeit und Feedback entsprechend ihrer Unterscheidung zwischen Aufgaben- und Prozessfeedback eine differenzierte Aussage getroffen werden. Auf Aufgabenebene ist ein geringer zeitlicher Abstand vorzuziehen, wenn Bearbeitung automatisiert werden soll. Bezieht sich das Feedback hingegen auf die Prozessebene, ist mittelbares Feedback effektiver, da Prozesse Zeit brauchen, um verarbeitet und reflektiert zu werden (Hattie & Timperley, 2007). Hierzu muss allerdings angemerkt werden, dass auch „delayed feedback“ in den Studien innerhalb eines Tages bedeutete (Bangert-Drowns, Kulik & Kulik, 1991) und somit keineswegs mit den Zeitabständen vergleichbar ist, die bei den Rückmeldeverfahren der Lernstandserhebungen (und auch anderer Schulleistungsmessung) entstehen. Zwar böten die Zeiträume ausreichend Zeit, um sich mit den Prozessen des Unterrichtens auseinanderzusetzen, für die jeweils gemessenen Bereiche gehen über den großen Zeitraum allerdings die situativen Anregungen verloren. Der Hinweis von Hattie und Timperley könnte eher als Anregung für ein Feedback an Schülerinnen und Schüler dienen oder aber in der zweiten Phase der Lehramtsausbildung zur Nachbesprechung von Unterrichtsbesuchen aufgegriffen werden.

An einer ausreichend differenzierten Aufgabenklassifikation fehlt es noch (Maier, 2009). Es kann statt einer systematischen Klassifikation hier nur eine unvollständige Liste wichtiger Aspekte gegeben werden. Bei der ausgeführten Tätigkeit kann zwischen der Aufgabenkomplexität und der Art der Aufgabe unterschieden werden. Für die Aufgabenkomplexität gilt, dass einfache kognitive Tätigkeiten eher nach einer Feedback-Intervention erfolgreicher ausgeführt werden als komplexere kognitive Aufgaben. Dies kann mit dem motivierenden Element von Feedback-Interventionen begründet werden. Bei kognitiv komplexen Aufgaben reicht eine höhere Motivation allein nicht aus, um eine Leistungssteigerung zu erreichen. Dort gelingt eine Leistungsänderung nur mittels systematischer Lernprozesse, d.h. bei angebotenen Verbesserungshinweisen (Kluger & DeNisi, 1996). Betrachtet man die Tätigkeiten in Experimenten oder Feldstudien in der Feedback-Forschung, muss man zu dem Schluss kommen, dass einfache kognitive Aufgaben, bei denen eine Leistungssteigerung durch Motivation erreicht werden kann, nur als mechanische Tätigkeiten existieren. Tätigkeiten aus Feldexperimenten sind hingegen stets komplexe Aufgaben, bei denen systematische Lernprozesse benötigt werden. Auch das Unterrichten muss als hochkomplexe Tätigkeit gesehen werden. Dies betrifft sowohl das Lehren im Klassenraum als auch die Vor- und Nachbereitung des Unterrichts, also auch die

Evaluation an sich. Bei der Art der Aufgabe ist entscheidend, ob Anstrengung oder Können im Vordergrund stehen (Coe, 1998). Die spiegelt sich auch in den beiden Dimensionen von zentralen Schulleistungsmessungen, Rechenschaftslegung und Qualitätsentwicklung, wider. Während hinter Rechenschaftslegung die Idee steht, Lehrkräfte mittels zusätzlicher Kontrolle zu einer besseren Leistung zu motivieren, dient Feedback als Mittel der Qualitätsentwicklung, um Verbesserungehinweise und Informationen zur Bewältigung der Aufgaben anzubieten. Wird Feedback vom Empfänger als Motivationsinstrument gesehen, steigert es insoweit die Leistung, wie benötigte Informationen verfügbar sind. Hier besteht bei unzureichender Informationslage die Gefahr der Überforderung und einem daraus resultierenden Leistungsabfall (Coe, 1998). Qualitätsentwicklung wird hingegen genau dann angeregt, wenn der Feedbackempfänger ein Bedürfnis nach zusätzlichen Informationen besitzt. Corbalan, Kester und van Merriënboer fanden in einem Experiment zumindest eine höhere Motivation derjenigen, die selbst wählen konnten, zu welcher Aufgabe sie ein Feedback erhielten (Corbalan, Kester & Merrienboer, 2009).

Für einen gewinnbringenden Umgang mit Feedback ist somit die Bedeutung der Aufgabe relevant (Kluger & DeNisi, 1996). Erkennt die gefeedbackte Person keinen oder nur einen geringen Bezug der Aufgabe zur eigenen Person, besteht eine geringere Motivation, sich mit den gegebenen Feedbackinformationen kritisch auseinanderzusetzen. In der Befragung von Lehrkräften zur ersten Durchführung von Lernstand8 (bzw. damals noch Lernstand9) fand Kühle die bereits beschriebenen unterschiedlichen Einschätzungen zwischen Lehrkräften verschiedener Schulformen, ob die gestellten Aufgaben für ihren Zweck angemessen seien (Kühle, 2010). Dem sollte und soll bei Lernstand8 mittels verschiedener Testheftversionen und dem Konzept der Standorttypen für den sozialen Vergleich vorbeugend begegnet werden.

Unter Situations- bzw. Persönlichkeitsvariablen sind Variablen wie die Klarheit, Spezifikation und der Anspruch der Ziele zusammengefasst (Maier, 2009). Allgemein verstehen Kluger und DeNisi darunter Zielstrukturen, die sich auf Ziele auf der Selbst-Ebene beziehen und handlungsleitend bei dem Umgang mit zurückgemeldeten Ist-Soll-Vergleichen wirken, Attributionsstile und das Selbstkonzept in seinen verschiedenen Formen (Ilgen & Davis, 2000). Auch Kontrollüberzeugungen oder die Tendenz zu Ängstlichkeit oder zu Altruismus zählen dazu (Coe, 1998). Vieles, welches für die Unterrichtsgestaltung relevant ist und im vorherigen Unterkapitel beschrieben wurde, lässt sich an dieser Stelle übertragen. Auf einige Aspekte muss allerdings noch einmal eingegangen werden.

An verschiedener Stelle findet man Hinweise darauf, dass ein hohes (Fähigkeits-) Selbstkonzept und vor allem eine hohe Selbstwirksamkeitserwartung als Moderatorvariable, die bei höherer Ausprägung auch eher negative Rückmeldungen akzeptieren lassen. Negatives Feedback wirkt hemmend, wenn die gefeedbackte Person nur ein geringes Selbstwertgefühl besitzt (Ditton & Arnoldt, 2004a; Hattie & Timperley, 2007; Maier, 2009).

Im Sinne des Konsistenzmotivs kann auch positives Feedback negative Auswirkungen haben, nämlich wenn Erfolg external attribuiert wird und Meta-Task Processes angeregt werden.

Zur Aufrechterhaltung des Selbstbilds ist es notwendig, gegen die zu positive Leistung Maßnahmen einzuleiten, die sich u.a. in reduzierter Motivation ausdrücken können (Kluger & DeNisi, 1996).

Es müssen möglicherweise außerdem Unterschiede zwischen allgemeiner SWE, Tätigkeits-SWE und Innovations-SWE bedacht werden: Eine Person mit hoher allgemeiner SWE erwartet möglicherweise mit größerer Wahrscheinlichkeit als eine Person mit gering ausgeprägter allgemeinen SWE eine positive Rückmeldung oder ein rückgemeldetes Defizit, welches gut behebbar ist, und ist daher eher bereit, sich einer Feedbacksituation zu stellen. Eine Person mit hoher Tätigkeits-Selbstwirksamkeit bzw. hohem Fähigkeitsselbstkonzept bzgl. der bewerteten Tätigkeit wird sich eher seltener einer Feedbacksituation aussetzen, da diese kein Weiterentwicklungspotenzial verspricht. Personen mit einer hohen Innovations-SWE sollten hingegen Feedbacksituationen suchen, weil sie zuversichtlich sind, etwaige Defizite systematisch beheben zu können.

Belebend soll auch eine vorherrschende Kommunikations- und Kooperationskultur auf die Auseinandersetzung mit den Rückmeldeangeboten wirken (Ditton & Arnoldt, 2004a; Gräsel, Fußangel & Pröbstel, 2006; Koch, 2011; Maier, 2009; Müller, 2010). Lehrerkooperation könnte externe Unterstützung ersetzen und flankieren, die für den Gebrauch solcher Rückmeldeangebote wie zentraler Lernstandserhebungen von Wissenschaftlern (Visscher & Coe, 2003) genauso gefordert werden wie von Lehrkräften und Schulleitungen (Schneewind, 2007a). Bestätigende Befunde für diese Annahmen liegen bisher aber noch nicht vor. Zwar gibt es einige Unterstützungsmaßnahmen wie umfangreiche Fortbildungskonzepte in Thüringen, diese führen aber nicht zwingend zu einer höheren Nutzung der Rückmeldeangebote (Maier, 2008a; Müller, 2010). Ditton und Arnold fanden in einer nicht repräsentativen Interventionsstudie zur Wirkung von Schülerfeedback immerhin, dass das Schülerfeedback von eng mit ihren Fachkollegen kooperierenden Lehrkräften positiver eingeschätzt wird. Dabei wird das Feedback möglicherweise weniger wegen als Informationsgewinn geschätzt, sondern eher als Form des sozialen Umgangs gesehen (Ditton & Arnoldt, 2004a). Krause, Stark und Mandl (2004) fanden in einem Experiment einen negativen Effekt von Kooperation auf die Wirkung von Feedback. Im Experiment zeigte sich, dass in Gruppen lernende Studierende weniger von Feedback zu einem (gemeinsam) bearbeiteten Test profitierten als individuell arbeitende Teilnehmer (Krause, Stark & Mandl, 2004). Dies ist auf die Fachgruppenkooperation aber nur bedingt übertragbar, weil sich die Rückmeldungen aus LSE als individuelle Rückmeldungen darstellen und lediglich in Kooperation verarbeitet werden sollen. Das berichtete Ergebnis der Studie beinhaltet außerdem auch keine Angabe über die Effektstärke.

Widersprüchlich sind die Befunde auch zum Zusammenhang von erlebter Belastung und Bereitschaft zur Teilnahme an Feedbackprozessen. Scherm hält Ergebnisse aus der Akzeptanzforschung aus der Wirtschaft für (bedingt) übertragbar, nach denen Personen Feedbackprozessen ablehnend gegenüberstünden, wenn sie besondere individuelle Belastung erwarteten (Scherm, 2002). Dies ist konform zu den Befunden über den

Zusammenhang von erlebter Belastung und Unterrichtsgestaltung. Ditton und Arnold (2004) fand in ihrer Befragung hingegen keinen Zusammenhang mit dem Merkmal der erlebten Belastung. Die Rücklaufquote war allerdings sehr gering, sodass besonders beanspruchte Lehrkräfte eventuell auch den Fragebogen gar nicht ausgefüllt haben.

Insgesamt lässt sich für die Situations- und Persönlichkeitsvariablen eine große Übereinstimmung mit dem Wirkungsmodell zur Unterrichtsqualität feststellen. Dies sollte nicht weiter verwundern: werden doch das Unterrichten und das Innovieren gleichermaßen als *Aufgabe* von Lehrkräften angesehen. Differenzen gibt es bei der detaillierteren Betrachtung des Selbstkonzepts und der Attribution von Leistung. Auch kommen mit den Facetten der Feedbackbotschaft und der Handlung zwei neue Variablenklassen dazu, die im vorherigen Modell nicht integriert sein konnten.

Abweichend von der Idee, einzelne Persönlichkeitsmerkmale zu identifizieren, gibt es auch noch die Vorstellung einer grundlegenden Innovationsbereitschaft. Diese Vorstellung kann zwar auch als eine Persönlichkeitsvariable aufgefasst werden (Ditton & Arnoldt, 2004), hat aber auch zu mehreren Typenmodellen geführt. Das älteste Typenmodell von Stamm charakterisiert den Evaluationsprozess ganzer Projektgruppen. Hosenfeld (2010), Groß Ophoff und Kollegen (2007, 2011) haben diesen Ansatz für VERA3 aber auch auf eine Individualtypisierung heruntergebrochen. Sowohl der projekt-basierende als auch der individuum-basierte Ansatz sollen folgend vorgestellt werden, da der Ansatz der Typisierung auf Individualebene ein paralleles Modell zum Typisierungsansatz des AVEM darstellt und somit einen Vergleich zwischen beiden Interpretationen der LSE erlaubt.

4.3.4 Innovationstypenansätze

Stamm dienen als Datengrundlage für ihr Typisierungsmodell Interviews mit sechsundfünfzig Projektbeteiligten von achtzehn Evaluationsprojekten aus dem Bereich des Bildungswesens[86]. Das Typenmodell beschreibt die Rezeption, den Transfer und den Nutzen von Evaluationsergebnissen mittels acht Vergleichsdimensionen: (1) Qualität der Evaluation (Sprache, Verständlichkeit, Glaubwürdigkeit, Arbeitsklima und Zufriedenheit mit der Zielsetzung der Evaluation), (2) Entscheidungskader/Akteure (persönliches Engagement, finanzielle Ressourcen, Vorerfahrungen mit Evaluation, Engagement der Beteiligten), (3) Evaluator/in (Commitment, zeitliche Ressourcen, Berufsorientierung, Reputation), (4) Konformität (Anwendbarkeit und Nützlichkeit), (5) Organisation (Art des Systems, Ausmaß der Veränderungsbereitschaft, Einbindung der Evaluation in die Gesamtstrategie), (6)

[86] Weitere Typisierungen bzw. Klassifizierungen zu zentralen Vergleichsarbeiten existieren beispielsweise von Hartung-Beck und Kuper (2009), Kuper und Hartung (2007) sowie von Diemer und Kuper (2011) auf Grundlage von Interviews. Da dieser Arbeit zwei quantitative Studien zugrunde liegen, beschränkt sich die Darstellung hier auf den umfangreichen Ansatz von Stamm. Es wird auf eine ausführliche Darstellung der anderen beiden Ansätze hier verzichtet.

Kontext und Interesse (Bildungspolitische Bedeutung, politische Brisanz, Initiierung, Widerstände, Unterstützung durch Vorgesetzte), (7) Dissemination und Diffusion (Umfang von Verbreitung der Ergebnisse) und (8) Rezeption, Transfer und Nutzen (direkt nachweisbare Veränderungsmaßnahmen) (Stamm, 2003). In den acht Vergleichsdimensionen wird deutlich, welche verschiedenen Aspekte zusätzlich zu denen, die schon bei einer Feedbackintervention betrachtet wurden, bei Evaluationsprojekten berücksichtigt werden müssen. Neben der Persönlichkeit der gefeedbackten Person berücksichtigt Stamm den Feedbackgeber als eigenständige Person und die Position der Institution und ihr Umfeld. Stamm liefert dadurch einen konkreten Vorschlag für die Situationsvariablen, die im vorherigen Abschnitt nur angedeutet, aber nicht konkret benannt wurden.

Die vier Typen im Modell von Stamm lassen sich wie folgt charakterisieren: „Typ 1 – Reaktion“ beinhaltet Evaluationen, die einerseits klar durch Überprüfung und Kontrolle veranlasst sind, gleichzeitig aber eine hohe Rezeption und Nutzung der Ergebnisse aufweisen. Die Evaluation wird hier als Reaktion auf Beschlüsse (wie beispielsweise Erlasse des zuständigen Ministeriums) umgesetzt und besitzt eine hohe bildungspolitische Bedeutung, beruht dadurch in der Regel aber auch auf professionellen Evaluatoren. Die Gefahr von Widerstand durch einzelne Akteursgruppen ist relativ hoch. Mit „Typ 2 – Innovation“ werden solche Evaluationen klassifiziert, die sich durch einen eindeutig entwicklungsorientierten Charakter auszeichnen. Evaluation ist hier als ein Bottom-up-Prozess zu verstehen. Die Qualität der Evaluation ist hoch, aber vor allem treten neben den hohen Veröffentlichungsgrad eine hohe Zufriedenheit mit dem Prozess und – besonders wichtig – eine hohe tatsächliche Nutzung der Ergebnisse. Der „Typ 3 – Blockade“ bezeichnet ebenfalls wie der Typ 1 Evaluationen, deren Ausgangspunkt neben der Umsetzung die Überprüfung und Kontrolle darstellen. Es kommt nur zu einem geringen Teil zu einem Nutzen der Ergebnisse und die meisten Beteiligten sind mit dem Prozess eher unzufrieden. Der hohe Widerstandsgrad kann sowohl externe Ursache (z.B. Ergebnisse für gefeedbackte Personen unverständlich, mangelhafte Rückmeldung zum Prozess durch Vorgesetzte) wie interne Ursachen (z.B. Handlungsalternativen unklar, Misstrauen in die Datenqualität, lediglich an externer Legitimation interessiert) als Hintergrund haben. „Typ 4 Alibi“ sammelt diejenigen Evaluationen zusammen, bei denen der legitimierende Charakter nicht nur als Hemmnis fungiert, sondern das wesentliche Ziel der Evaluation darstellt. Der Nutzen der Ergebnisse wird nur scheinbar angestrebt. In den meisten Fällen handelt es sich hier um Top-down-Prozesse. Die Folge sind eine geringe Nutzenquote und Zufriedenheit (Stamm, 2003). Bezogen auf die Lernstandserhebungen kann ihr Vier-Typenmodell auf zweierlei Art verstanden werden: Lernstandserhebungen können in der konkreten Manifestierung wie Lernstand8 als ein Projekt aufgefasst und entsprechen einem der vier Typen zugeordnet werden. Genauso ist es aber möglich, den jeweiligen Umgang der einzelnen Schulen als jeweils eine eigene Evaluation zu betrachten und die einzelnen Schulen einzustufen. Besonders für die Schulebene kann das Modell eine Problemanalyse unterstützen, wenn die Charakterzüge der vier Typen als wahrgenommene Charakteristika verstanden werden. Aus

der Einordnung von Schulen ließe sich der Erfolg einer Maßnahme wie die zentralen Vergleichsarbeiten messen. Um allerdings den vollen Umfang zu erfassen und konkrete positive wie negative Folgen einer solchen Maßnahme einordnen zu können, muss neben der von Stamm hauptsächlich in den Blick genommenen Prozessebene auch die Unterrichtsebene analysiert werden. Folgen für die Unterrichtsebene ergeben sich aber aufgrund der Klassenraumautonomie der Lehrkraft durch die Beteiligung an Evaluationsmaßnahmen auf individueller Ebene.

Eine Typenklassifikation von Rezeptionstypen auf Individualebene hat Hosenfeld in ihrer Interventionsstudie zu videogestützten Unterrichtsrückmeldungen mittels latenter Klassenanalysen berechnet (Hosenfeld, 2010). Hosenfeld unterscheidet zwischen Rezeption, Reflexion und Aktion (geplante Veränderungen) jeweils für die schriftlichen Rückmeldungen und die Videorückmeldungen. Über die sechs Kategorien stellte sich eine Drei-Klassen-Lösung als die statistisch sinnvollste heraus (über die Informationskriterien AIC[87], BIC und CAIC sowie über die Datenpassung mittels Pearson-χ^2-Wert und Cressie-Read-Statistik ermittelt) im Vergleich zu den ebenfalls geprüften Ein-, Zwei- und Vier-Klassen-Lösungen. Die Wahrscheinlichkeit für eine richtige Klassenzuordnung liegt bei 94% (Typ 1 und Typ 2) bzw. 99% (Typ 3). Insgesamt haben an dieser Studie mit n= 46 wesentlich weniger Lehrkräfte teilgenommen als an der Befragung von Groß Ophoff u.a., sodass die nachfolgend berichteten Ergebnisse nur erste Anhaltspunkte liefern können.

Der Typ 1 in der Klassifikation von Hosenfeld umfasst 59% aller Untersuchungsteilnehmer und zeichnet sich über beide Rückmeldearten durch eine selbstberichtete starke Reflexion aus. Auch die Skalenwerte für die Video-Aktion-Kategorie sind relativ hoch und die Werte für die Rezeption und Aktion der schriftlichen Rückmeldungen sind jeweils die höchsten aller drei Typen. Hier eingruppierte Lehrkräfte erreichen in der Selbsteinschätzungs-Skala „Kenntnis des eigenen Unterrichtsstils“ den deutlich niedrigsten Wert.

Den Typ 2 bildet 24% aller Untersuchungsteilnehmer. Charakteristisch sind deutlich abfallende Werte von der Rezeption über die Reflexion bis zur Aktion bei beiden Rückmeldearten. Für die schriftliche Rückmeldung sind die Werte insgesamt aber höher, für die Rezeption ist der Wert hier ähnlich hoch wie bei Typ 1. Hosenfeld erklärt die Differenz zwischen einerseits wahrgenommenen hohen in die Rezeption investierter Zeit und geringer Reflexions- und Aktionshandlungen über den geringen wahrgenommenen Handlungsbedarf nach Sichtung der zurückgemeldeten Ergebnisse. Dazu passend ist der signifikant höhere Skalenwert in der Selbsteinschätzungs-Skala „Kenntnis des eigenen Unterrichtsstils“ im Vergleich zu den beiden anderen Typen. Auch die Werte in den Kategorien „Berufszufriedenheit“, „Reflexionsbereitschaft“, „Selbstwirksamkeit“ sowie „Kooperationsbereitschaft“ sind (aber nicht signifikant) höher.

[87] Akaike Informationskriterium (AIC), Bayes‘sches Informationskriterium (BIC) und Constistent Akaike Informationskriterium (CAIC)

Dem Typ 3 werden 18% der Untersuchungsteilnehmer zugeordnet. Lehrkräfte dieses Typs verhalten sich im Bereich des Videofeedbacks konträr zu Typ 2, zeigen in eigener Wahrnehmung folglich eine geringe Rezeption, aber hohe Reflexion und Handlungsbereitschaft. Schriftliche Rückmeldungen werden hingegen weder rezipiert noch reflektiert und lösen somit auch keine geplanten Veränderungen aus. Lehrkräfte dieses Typs zeichnen sich darüber hinaus durch eine hohe Reflexionsbereitschaft im Vergleich zu Lehrkräften des Typ 1 aus.

Groß Ophoff, Hosenfeld und Koch (2007) bzw. Groß Ophoff (2013) haben Analysen zum Rezeptions- bzw. Reflexionsverhalten[88] von Lehrkräften im Zusammenhang mit VERA3 durchgeführt. In die Typenklassifikation wurden Lehrkräfte aufgenommen, die Fragen zur Verständlichkeit der Ergebnisrückmeldung von VERA3 2004 bis 2008, zur Intensität der Auseinandersetzung mit den Ergebnissen und zur erlebten Nützlichkeit der zentralen Rückmeldeelemente (Fähigkeitsniveau-Verteilung auf Schüler-, Klassen- und Landesebene sowie im „Fairen Vergleich“) beantwortet hatten. Auch für Lernstand9 hat Groß Ophoff die Analyse auf der Grundlage von Befragungsdaten aus dem Schuljahr 2004/05 durchgeführt. Es zeigte sich über alle Stichproben hinweg einzelne Reflexionstypen wiederkehrend, aber nicht konsistent alle Typen in allen Analysen. Auch wies in den einzelnen Datensätzen mal eine Vier-Klassen-Lösung (mit Restklasse) und mal eine Fünf-Klassen-Lösung (mit Restklassen) die besten statistischen Kennwerte auf. Insgesamt zeigte sich hingegen für die Kombination der Datensätze zu VERA3 eine Sechs-Klassen-Lösung (mit Restklasse) als die annehmbarste (Groß Ophoff, 2013). Die inhaltlich interpretierbaren Typen unterschieden sich dabei sowohl darin, ob alle vier Verteilungsniveaus in gleicherweise bewertet wurden oder ob ein Gefälle von Schülerebene zu „Fairen Vergleichen“ bestand, als auch darin, wie stark eine Auseinandersetzung mit den Ergebnissen und ihre Nutzung stattfand (Groß Ophoff, 2013; Groß Ophoff, Hosenfeld & Koch, 2007). Die Befunde von Groß Ophoff und Kollegen zeigen vor allem die Anfälligkeit des Vorgehens bei Typisierungen und lassen sich inhaltlich zumindest nicht außerhalb von Trendanalysen sinnvoll erklären.

Ebenfalls auf Individualebene haben Ditton und Arnoldt eine Klassifizierung mittels einer Clusteranalyse vorgenommen. Sie teilten einundneunzig Lehrkräfte, die an QuaSSU teilnahmen, aufgrund der Bewertung der Form bzw. Verständlichkeit der Rückmeldung, der Kompetenz der Schülerinnen und Schüler als Feedbackgeber und der Verwendbarkeit der Rückmeldungen in zwei Cluster ein (Ditton & Arnoldt, 2004). Neunundsechzig Lehrkräfte wurden dem ersten Cluster zugeordnet, welches sich dadurch auszeichnete, dass die Rückmeldung als verständlicher, die Schülerinnen und Schüler als kompetent genug und die Ergebnisse als verwendbar eingeschätzt wird. Dem zweiten Cluster wurden dreiundzwanzig Lehrkräfte zugeordnet und diese bescheinigten zwar eine etwas höhere Verständlichkeit, sprachen den Schülerinnen und Schülern aber eher die Kompetenz zur Beurteilung des Unterrichts ab und hielt die Rückmeldungen tendenziell für nutzlos. Unter den „Kritikern“, wie das zweite Cluster benannt wurde“ sind männliche Lehrkräfte und Überfünfzigjährige

[88] Ursprünglich haben Groß Ophoff, Hosenfeld und Koch (2007) von Rezeptionstypen gesprochen. Groß Ophoff (2013) hält Reflexionstypen mit Verweis auf die Formulierung der verwendeten Items für angemessener.

überrepräsentiert. Auch finden sich mehr Mathematiker als Deutsch- und Englischlehrkräfte im zweiten Cluster als es dem tatsächlichen Verhältnis in der Stichprobe entspricht.

4.3.5 Befunde zur Rezeption von Ergebnissen aus Schulleistungsmessungen

Der von Hosenfeld sowie von Ditton und Arnoldt gewählte Zugang ist charakteristisch für die Rezeptionsforschung. Vorliegende Rezeptionsforschung rekurriert häufig auf die FIT von Kluger und DeNisi (Kluger & DeNisi, 1996) und vermehrt auch auf die Handlungstheorie von Frese und Zapf (Frese & Zapf, 1994) (vgl. beispielsweise neben Groß Ophoff u.a. 2007 sowie Ditton & Arnoldt 2004; Maier, 2009; Müller, 2010; Schneewind, 2007a[89]). Der Zugang der Rezeptionsforschung muss aber als problemorientiert bezeichnet werden, während traditionelle Feedbackforschung experimentell oder quasi-experimentell vorgeht. Daraus resultiert u.a. eine introspektive Perspektive der Rezeptionsforschung (Dedering, 2011) im Vergleich zur Beobachterperspektive der traditionellen Feedbackforschung. Statt die Wirkung von Eigenschaften der Feedbackbotschaft, der Aufgabe, der Person oder der Situation des Feedbackkontexts erklären zu wollen, werden Zusammenhänge des durch die Rezipienten wahrgenommenen Feedbackkontextes mit potenziellen Veränderungs-Handlungen von Akteuren aus dem Schulbereich (Schulleitung, Bildungsverwaltung, Erziehungsberechtigte und besonders Lehrkräfte) untersucht. Durch dieses Vorgehen nicht ausreichend erforschte Leerstellen, beispielsweise darf nicht nur wahrgenommene Verständlichkeit mit tatsächlichem Verständnis verwechselt werden (Müller, 2010), auch könnten Gründe wie zu hohe Belastung für die Ablehnung von Evaluationsverfahren nur vorgeschoben (Kohler & Schrader, 2004) oder die tatsächlichen Gründe den befragten Lehrkräften selbst nicht bewusst sein. Auch die selbstberichteten Unterrichtsveränderungen müssen hinterfragt werden, da diese nicht substanziell und in der beabsichtigten Weise ausfallen müssen (Clausen, 2002; Kohler & Schrader, 2004). Einzig Hosenfeld weicht neben Koch (Koch, 2011) mit ihrer Interventionsstudie zur Wirkung von Videofeedback von diesem charakteristischen Vorgehen ab und entspricht durch die Kontrolle der Lese- und Mathematikleistung zu zwei Messzeitpunkten dem üblichen Vorgehen in der Feedbackforschung. Substanzielle Effekte der angebotenen Rückmeldungen auf die Leistung der Schüler und Schülerinnen können sich aber nicht nachweisen (Hosenfeld, 2010).

Anders als die drei Typisierungsansätze (Hosenfeld 2010; Groß Ophoff u.a. 2007, Ditton & Arnoldt 2004) sind die meisten anderen Studien rein deskriptiver Natur und berichten lediglich Globalergebnisse. Exemplarisch seien hier drei Studien mit ihren besonders bemerkenswerten Ergebnissen skizziert, um die Vorgehensweisen und charakteristischen Fragestellungen und Ergebnisse darzustellen.

[89] Für einen breiteren Überblick siehe Dedering (2011).

Maier untersuchte die Einführung zentraler Vergleichsarbeiten in der Sekundarstufe I in Baden-Württemberg und verglich Akzeptanz und wahrgenommene Nützlichkeit mit denen der Thüringer Kompetenztests (hier: Kompetenztestes in Klasse 6 und 8). Dazu hat Maier teilnehmende Lehrkräfte aus Baden-Württemberg von 2004 bis 2007 mittels Fragebögen und 2006 mittels Interviews befragt, wobei von 180 Untersuchungsteilnehmern von 2005 bis 2007 längsschnittlich erhobene Daten vorliegen. Unter Lehrkräften aus Thüringen 2007 hat Maier ebenfalls Fragebogen- und Interviewdaten erhoben (Maier, 2009). Fragebögen und Interviews beinhalteten Fragen über die allgemeine Akzeptanz zentraler Tests, nach ihrer Einschätzung zur curricularen Validität der zentralen Vergleichsarbeiten und nach der erlebten Beanspruchung durch die Tests. Die Nützlichkeit wurde in diagnostische Hinweise, Hinweise zur Notengebung, Hinweise für zukünftige Wiederholungsphasen und inhaltliche Änderungen unterschieden. Bemerkenswert sind an Maiers Ergebnissen eine abnehmende Akzeptanz der zentralen Vergleichsarbeiten und der ihnen zugesprochene Nutzen in Baden-Württemberg, aber auch in Thüringen. Maier kommt in einer vergleichenden Betrachtung zu der Einschränkung, dass der Umfang der beabsichtigten und durchgeführten Nutzung genauso wie der wahrgenommene Nutzen und die Akzeptanz von zentralen Vergleichsarbeiten in einem mit Unterstützungsmaßnahmen flankierten System höher sind (Maier, 2008b). Zusätzlich wurde die Lehrerselbstwirksamkeit erhoben, die aber entgegen der FIT keine signifikante Wirkung auf die Akzeptanz oder Nützlichkeitseinschätzung zeigte. Maier deutet dies derart, dass zwischen Nutzungsmöglichkeiten und Nutzung unterschieden werden muss und die SWE erst bei der tatsächlichen Nutzung eine moderierende Wirkung besitzt (Maier, 2008a).

Schneewind befragte zu fünf Messzeitpunkten 52 teilnehmende Lehrkräfte des Projekts BeLesen schriftlich und zehn weibliche Lehrkräfte zusätzlich in Interviews zu Einstellungen zu zentralen Tests und deren Ergebnisrückmeldungen, zur Verständlichkeit der angebotenen Rückmeldung, die Verwendung der Ergebnisse und Erklärungen für die zurückgemeldeten Testergebnisse. Hierbei zeigte sich neben der schon in 4.3.1 beschriebenen Fehlattribution vereinzelte Nutzung der Ergebnisse zu diagnostischen Zwecken, vor allem aber gaben die Lehrkräfte mehrheitlich an, zwar an Informationen über den Lernstand ihrer Schüler und Schülerinnen interessiert zu sein, die Informationsbeschaffung sollten aber keine zusätzliche Anstrengung bedingen und nicht zu höheren Ansprüchen anderer bzw. zu Kritik führen. Außerdem fehlte es den Lehrkräften mehrheitlich an Wissen über Handlungsalternativen, um eine Unterrichtsentwicklung überhaupt umsetzen zu können (Schneewind, 2007, 2007a).

Müller befragte 40 Grundschullehrkräfte mittels Fragebögen über den Umgang mit den Ergebnisrückmeldungen im Rahmen des Projekts KOALA-S. Es wurden neben Lehrermerkmalen auch Unterrichts- und Schulmerkmale erhoben. Dabei zeigten sich die Lehrervariablen als bedeutsamer im Vergleich zu den Schul- und Unterrichtsvariablen. Die angebotenen Rückmeldungen berücksichtigten dabei den in Schneewinds Studie geäußerten Wunsch der Lehrkräfte nach einfachen Darstellungen. Diese wurden auch durchaus als verständlich und nachvollziehbar eingeschätzt. Auch gaben die Lehrkräfte an, in diesen einfachen Darstellungen ein Nutzungspotenzial für Unterrichtsreflexionen zu erkennen.

Weiter bestätigte sich Maiers Befund, dass die SWE keinen Einfluss auf die Einschätzungen über das bereitgestellte Instrument aufwies. Interessant ist auch der gefundene Zusammenhang zwischen Umweltorientierung bei der Attribution von Schülerleistungen und dem Ausmaß an als notwendig betrachteten Maßnahmen. Je mehr Lehrkräfte Umweltbedingungen als leistungsrelevant ausmachten, desto größer war nach Erhalt der Rückmeldungen aus ihrer Sicht die Notwendigkeit, Maßnahmen zur Verbesserung des schulischen Qualitätsmanagements zu ergreifen (Müller, 2010).

In einem Überblick zu Rezeptionsstudien hat Dedering alle externen Leistungserhebungen berücksichtigt, die sich auf fachliche und überfachliche Schülerleistungen beziehen und auf die Rezeption durch die beteiligten Lehrkräfte rekurrieren. Dies umfasst sowohl kontinuierliche Lernstandserhebungen wie VERA3 und VERA8 als auch einmalige Leistungsvergleichsstudien einzelner Bundesländer wie LAU sowie internationale Leistungsvergleichsstudien wie PIRLS, PISA und DESI und Interventionsstudien aus Berlin, der Schweiz und Österreich. Für den Bereich Rezeption und Reflexion bilanziert Dedering u.a.: (a) Die wahrgenommene Verständlichkeit fällt über alle betrachteten Studien eher positiv aus. Die Mehrheit der Befragten hält die Rückmeldungen zumindest in einzelnen Komponenten für verständlich und Schwierigkeiten scheinen bei Durchführungswiederholungen abzunehmen. (b) Allgemein schätzen die Befragten die Nützlichkeit und Bedeutsamkeit positiv ein. Der Grad der Nützlichkeit scheint aber stark von Komplexität der Rückmeldung und somit von Verständnis der angebotenen Rückmeldungen abzuhängen. Konkret für die Unterrichtsentwicklung wird die Nützlichkeit weniger hoch eingeschätzt als zur Diagnose, zur Selektion, zur Überprüfung des Lehrplans und andere Nutzungsmöglichkeiten, die keinen direkten Bezug zu eigenen Handlungen aufweisen (vgl. auch Bonsen et al., 2006; Kühle & Peek, 2007). (c) Die Bereitschaft zur Auseinandersetzung mit den Ergebnisrückmeldungen ist bei Lehrkräften eher mäßig und nur bei Schulleitungen hoch. Die eingeschätzte Nützlichkeit spielt hier erwartungskonform eine Rolle. (d) Eine tatsächliche Auseinandersetzung scheint auf individueller Ebene allerdings durchaus stattzufinden, wohingegen auf kooperativer Ebene dies nur für den Austausch zwischen Fachlehrern innerhalb der Jahrgangsstufe berichtet wird. (e) Der Fokus der Auseinandersetzung liegt grundsätzlich auf den Leistungsdaten und den angebotenen Vergleichswerten, weniger auf kriterialen Aufgabenanalysen. Ob dabei das Gesamtergebnis mit dem erwarteten Ergebnis, das Klassenergebnis mit den Ergebnissen von Parallelklassen oder individuelle Schülerergebnisse in den Blick genommen werden, variiert allerdings zwischen den Studien.

Für die im Helmke-Modell als „Aktion“ bezeichneten Folgemaßnahmen kommt Dedering zu dem Schluss, dass die Rückmeldeergebnisse eher selten zu einer (berichteten) Verhaltensänderung bei Lehrkräften (und Schulleitungen) führten. Wurden Veränderungen berichtet, handelte es sich vorwiegend um eingefügte Wiederholungsmaßnahmen und Vertiefungen von Stoffgebieten, in denen erwartungswidrige oder schlechte Ergebnisse erreicht worden waren. Weniger häufig wurden die Reflexion über Unterrichtsmethoden, Unterrichtsziele und Leistungsdifferenzierungen berichtet. Ebenfalls eher selten gaben die Lehrkräfte an, häufiger mit Kollegen kooperieren zu wollen, indem beispielsweise

didaktische und methodische Absprachen getroffen oder Aufgaben und Materialien ausgetauscht werden (Dedering, 2011).

Für den Wirkungszusammenhang fasst Dedering zusammen, dass die Verarbeitungsprozesse durch die Akzeptanz von externer Leistungsmessung und ihrer wahrgenommenen Nützlichkeit moderiert werden. Dies hatte sich schon in der Fragebogenerhebung zu Lernstand9 gezeigt, die Bonsen, Büchter und Peek 2004 durchgeführt haben (Bonsen et al., 2006). Gleiches gelte, so Dedering, für die Selbstwirksamkeitserwartung und das pädagogische Interesse sowie für professionelle Überzeugungen über die generelle Verwendbarkeit von Schulleistungsdaten (Dedering, 2011). Zumindest für die Selbstwirksamkeitserwartung bleibt mit den Befunden von Maier (Maier, 2008a) und (Müller, 2010) allerdings ein Fragezeichen, da Rezeptionsstudien eben genau nicht die tatsächlich durchgeführten Unterrichtsänderungen abbilden können. Insgesamt findet Dedering eine positive Wirkung der Ergebnisrückmeldungen maximal in Ansätzen belegt (Dedering, 2011). Die vielfach geäußerte Hoffnung, durch Ergebnisrückmeldungen innerschulische Diskussionsprozesse anzuregen (Kohler & Schrader, 2004; Visscher, 2008), hat sich folglich nicht erfüllt.[90]

[90] Zu dieser Einschätzung kommt auch Schneewind (2007a).

4.4 Das erweiterte Lehrer-Handlungskompetenzmodell

In (4.2) wurden die verschiedenen Bereiche eines Lehrer-Handlungskompetenzmodells für den Unterricht dargestellt und diskutiert. Als Ausgangspunkt diente das für COACTIV entwickelte Handlungskompetenzmodell mit den Bereichen Wissen/Können, Überzeugungen, motivationale Orientierung und Selbstregulation und zum Teil auch das Modell zum Projekt TEDS-M bzw. MT21. Die Bedeutung der vier Bereiche des COACTIV-Modells konnte nachvollzogen werden, aber die ebenfalls dargelegte Kritik an dieser Aufteilung legt eine Neueinteilung und Neubezeichnung nahe.

Es wurde außerdem in (4.2.1) angesprochen, dass auch ein auf das Unterrichten ausgelegtes Lehrer-Handlungskompetenzmodell eine Innovationskompetenz einschließen muss (Altrichter, 2000; Kiper, 2009; Schelten, 2009). Durch die vier definierten Kompetenzbereiche für Lehrkräfte durch die KMK „Unterrichten“, Erziehen“, Beurteilen“ und „Innovieren“ (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2004) wird diese These ebenfalls unterstützt. Diese Innovationskompetenz betrifft nicht nur langfristige und einmalige Reformvorhaben in einer Schule, wie sie in der Mitarbeit in Steuer- und Projektgruppen, in der Arbeit am Schulprogramm und -profil, bei der Organisation und Durchführung von internen Fortbildungen und durch die Veränderung der schulischen Organisationsstruktur ausgedrückt wird. Die Evaluation und Weiterentwicklung des ganz individuellen Unterrichts gehört genauso zu den alltäglichen Aufgaben des Lehrers wie das gemeine Lehren und Lernen Bestandteil von Schule ist. Auch die Weiterentwicklung und Evaluation des eigenen Unterrichts stellen eine Herausforderung dar, die Ressourcen aller vier Kompetenzbereiche voraussetzen. Ähnlich argumentieren auch Hense und Mandl in ihrem Plädoyer für Evaluations- und Selbstevaluationskompetenz für Lehrkräfte. Die von ihnen zugrunde gelegte „Anforderungsprofile an Evaluatorinnen und Evaluatoren“ der Gesellschaft für Evaluation beinhalten neben Methodenwissen Organisations-, Vermittlungs- und Kooperationskompetenz auch Praxiserfahrung und Wissen über die Theorie und Ideengeschichte der Evaluation (Hense & Mandl, 2009). Für Tenorth zählt auch eine bildungstheoretische Reflexion zum Innovationsanspruch von Lehrkräften. Lehrkräfte müssten entscheiden (und entscheiden können), welche Schlüsse sie aus den wissenschaftlichen Theorien und Fremdbeschreibungen ziehen (Tenorth, 2006).

Im Folgenden wird daher ein Lehrer-Handlungskompetenzmodell skizziert, welches die voran dargelegten Diskussionen des COACTIV-Modells berücksichtigt und Altrichter, Hense, Kiper, Mandl und Schelten folgend die Innovationskompetenz in dieses Modell integriert

In Analogie zum COACTIV-Modell teilt sich das hier postulierte Modell in die vier Kompetenzbereiche (1) Wissen und Können, (2) Kausal- und Zielüberzeugungen, (3) personenbezogene Überzeugungen des Berufserlebens und (4) Kompetenz- und Kontrollüberzeugungen (vgl. Abb. 4.3).

Abbildung 4.3 Das erweiterte Modell der Lehrer-Handlungskompetenz

4.4.1 Wissen und Können im erweiterten Lehrer-Handlungskompetenzmodell

Neben dem fachlichen, fachdidaktischen und allgemein-pädagogischen Wissen, dem Beratungswissen und dem Organisationswissen bedarf es Innovationswissen. Um den Lehr- bzw. Lernerfolg angemessen diagnostizieren und weiterführende Unterrichtsentwicklung durchführen zu können, muss auf fachliches und fachdidaktisches Wissen zurückgegriffen werden können. Darüber hinaus muss eine Lehrkraft über die von Kiper (2009) genannten Fähigkeiten im Umgang mit Daten verfügen. Mit speziellem Blick auf datengestützte Schulentwicklung listet Kiper das methodische, Diagnose- und theoretisches Wissen zu Erhebung, Auswertung und Interpretation von Daten, Wissen über Problemlöseprozesse und Managementfähigkeiten und verknüpft dies mit Beratungs- und Organisationswissen (s.4.2.1.5) (Kiper, 2009). Lehrer müssen wissen, wie Wissen über den Lernstand von Schülerinnen und Schülern erhoben werden können und wie bereits vorliegende Daten (beispielsweise aus Lernstandserhebungen) interpretiert werden müssen. Und schließlich sind Kenntnisse über den Zweck und die Funktionsweise der verschiedenen Instrumente der Schulleistungsmessung notwendig, um deren verschiedenartige Ergebnisse abgemessen einschätzen zu können und zu nutzen (oder eben wie bei international vergleichenden Schulleistungsstudien nicht zu nutzen).

An diesem Wissen scheint es Lehrkräften in Teilen aber zu mangeln. Müller und Kollegen konnten diesbzgl. in zwei Studien unter Lehramtsstudierenden zeigen, dass es Lehramtsstudierenden häufig nicht gelingt, ausreichende Statistikkenntnisse bis zum Ende ihres Studiums zu erwerben, um die übliche deskriptiven Statistiken zu verstehen, die bei zentralen Vergleichsarbeiten in Bayern zurückgemeldet werden. Die Lehramtsstudierenden wurden gebeten, die auf Klassenebene mit Vergleichsgruppe dargestellten Ergebnisse vierer Unterrichtsfächer nach Handlungsdringlichkeit zu sortieren. Dazu mussten Mittelwerte und Stichprobenfehler gleichzeitig analysiert werden. Insbesondere der Umgang mit den Stichprobenfehlern als Anhaltspunkt für die Streuung des tatsächlichen Mittelwerts bereitete dem Großteil der Untersuchungsteilnehmer große Schwierigkeiten (Müller, 2010).

Ähnliches konnten Müller und Hahn (2011) in einer Experimentalstudie replizieren, die unter Essener Lehramtsstudierenden durchgeführt wurde. Auch konnten die Untersuchungsteilnehmenden mehrheitlich zu drei Verfahren der Schulleistungsmessung (internationale Schulleistungsvergleiche, zentrale Vergleichsarbeiten und zentrale Abschlussprüfungen) nicht diejenigen Ziele (Überblick über den Bildungsstand in der Bundesrepublik, Hinweise für die Verbesserung des Unterrichts, Hinweise für die Gestaltung von Klassenarbeiten, Rückmeldung über den individuellen Kompetenzstand der Schülerinnen und Schüler, Prognose für den zukünftigen individuellen Bildungserfolg der Schüler und Schülerinnen, Identifikation guter und schlechter Schulen) identifizieren, die sich mit den jeweiligen Verfahren jeweils erreichen lassen (Müller & Hahn, 2011).

Auch Hahn (2008) stellte in einer qualitativen Interviewstudie mit siebzehn Lehrkräften aus dem Raum Dortmund bei einigen Lehrkräften mangelnde Kenntnisse über die Verwendungsmöglichkeiten von Vergleichsarbeiten und die Unterschiede zu zentralen Prüfungen wie „ZP 10“ und dem Zentralabitur fest. Nur einigen Lehrkräften war bewusst, dass die Ergebnisse aus Lernstand8 in erster Linie eine Rückmeldung über den Unterrichtserfolg darstellen sollen und sich dementsprechend an die Lehrkräfte selbst richten, nur in zweiter Linie an die Schülerinnen und Schüler (Hahn, 2008).

Koch zeigte in einem Experiment, wie sich die Rezeptionsart von Lehrkräften ändert, wenn sie über bessere statistische Kenntnisse verfügen. Lehrkräfte tendieren häufig dazu, Ergebnisse aus zentralen Vergleichsarbeiten lediglich auf individueller Schülerebene zu betrachten und zu attribuieren. Dies änderte sich, wenn sie die statistischen Herausforderungen von Rückmeldungen in einer Fortbildung trainierten (Koch, 2011).

4.4.2 Kausal- und Zielüberzeugungen im erweiterten Lehrer-Handlungskompetenzmodell

Gegenstandsbezogene Überzeugungen umfassen auch Kausalüberzeugungen und Zielüberzeugungen. Zielüberzeugungen beinhalten wiederum Wertvorstellungen und persönliche Ziele, Entwicklungsziele wie auch Lernziele und Erziehungsziele.

Kausalüberzeugungen bilden ein paralleles Konstrukt zu den Wissensdomänen und beziehen sich auf denselben Anwendungsbereich (die Fachwissenschaft, die Fachdidaktik, pädagogische, psychologische und sozialogische Fragestellungen). Kausalüberzeugungen sind in ihrer Handlungswirksamkeit von Wissenskognitionen nicht zu unterscheiden, generieren sich aber (standarisierten) Lehrmedien, sondern generieren sich aus Erfahrungs-Lerngelegenheiten. Sie unterliegen dadurch einer starken Interpretation und sind besonders widerspruchsanfällig in Verbindung zu anderen Kausalüberzeugungen.

Bzgl. der Nutzung von Evaluationsgelegenheiten sind vor allem Kausalüberzeugungen über die unterstellten Funktionen der zur Evaluation eingesetzten Instrumente und über die Einflussmöglichkeiten des Unterrichtsgeschehens auf die Lernleistung von Schülerinnen und Schülern entscheidend. Die Größe des Anteils am Lernerfolg, den Lehrkräfte ihrer eigenen Person bzw. ihrem eigenen Handeln zuschreiben, bedingt das Ausmaß der Motivation für Lehrkräfte, sich überhaupt datengestützter Schulentwicklung zu befassen. Sehen sich Lehrkräfte nicht für den Lernerfolg verantwortlich, fehlt es dazu an Anreizen. Unterrichtsentwicklung kann nur dann reizvoll sein, wenn sie darauf abzielt, persönliche Ressourcen zu erhalten oder zu steigern.

Weiter spielen auch hier die Überzeugungen eine Rolle, die abstraktes Unterrichten (ohne Innovationskomponente) bedingen. Innovationsprozesse bedeuten das Unterrichtshandeln zu reflektieren, welches dieselben Überzeugungen als Prämissen verwendet, die auch schon handlungsleitend bei der Planung und Durchführung des Unterrichts waren.

Schließlich braucht ein erfolgreicher Innovationsprozess ein bestimmtes professionelles Selbstverständnis (Bauer, 2009). Lehrkräfte müssen die in den Standards der KMK formulierten Innovationsaufgaben als Teil ihres Selbstverständnisses begreifen, um an Innovationsprozessen angemessen teilzunehmen und keine Form des von Steins charakterisierten Widerstand zu leisten (Steins, 2009).

4.4.3 Berufserleben als personenbezogene Überzeugungen im erweiterten Lehrer-Handlungskompetenzmodell

Zu den personenbezogenen Überzeugungen (im beruflichen Kontext) gehören Fähigkeitselemente (Widerstandsfähigkeit) sowie motivationale (Arbeitsengagement) und emotionale Elemente (Arbeitszufriedenheit, erlebte berufliche Beanspruchung).

Die Bedeutung der personenbezogenen Überzeugungen ergibt sich nach Altrichter dadurch, dass Innovationsprozesse zeitliche Ressourcen, aber andere Rahmenbedingungen wie besser ausgestattete Arbeitsplätze, Teamarbeitsplätze und eigene Budgets für Lehrerteams erfordern (Altrichter, 2000). Daraus lässt sich schließen, dass eine Innovationskompetenz motivationale Orientierung und Selbstregulationsfähigkeiten erfordert. Innovationsprozesse

sind Teil des Arbeitsumfangs und können nur in diesem Rahmen erfolgreich durchlaufen werden, wenn ausreichend Ressourcen zur Verfügung stehen, um einerseits in Teamarbeit Evaluationen durchzuführen um aber andererseits auch die kognitive und emotionale Leistung erbringen zu können, die Veränderungen des eigenen Handelns und der eigenen Gewohnheiten, der eigenen Ziele und Skripte erfordern. Unterrichten kann als eine Art Feedbackphase gesehen werden, weil Lehrkräfte durch die Schülerinnen und Schüler im Unterricht dauerhaftem Feedback ausgesetzt sind, das sich durch Schülerantworten und die Teilnahme der Schüler und Schülerinnen am Unterrichtsgeschehen manifestiert.

Auch hier sind Kausalüberzeugungen entscheidend, nämlich Überzeugungen darüber, in welcher Weise persönliche Ressourcen und in welcher Weise externe berufliche Ressourcen Handlungsoptionen im Unterrichtsgeschehen bedingen.

4.4.4 Kompetenz- und Kontrollüberzeugungen im erweiterten Lehrer-Handlungskompetenzmodell

Als Kompetenz- und Kontrollüberzeug muss in erster Linie die Selbstwirksamkeitserwartung genannt werden. Bisher liegen nur für diese Kompetenz- und Kontrollüberzeugung Studien vor, die einen direkten und vermittelten Zusammenhang zu anderen unterrichtsrelevanten Größen zeigen. Auch fachspezifische Fähigkeitsselbstkonzepte wären eine hierzu verortende Überzeugung.

Die Schnittmenge mit der Innovationskompetenz stellt sich hier wie über die doppelte Unsicherheit dar: Lehrkräften muss die Überzeugung zu Eigen sein, den Prozess erfolgreich durchleben zu können. Bei ihnen muss zweitens die Überzeugung vorhanden sein, die als besseren Handlungen und Ziele entdeckten Ergebnisse dieses Prozesses auch umsetzen zu können.

Empirischer Teil

Dem empirischen Teil liegt ein Projekt mit zwei parallelen quantitativen Studien zugrunde. Beide Parallelstudien waren als schriftliche Befragungen konzipiert[91] und wurden gleichzeitig im April/Mai 2010 durchgeführt. Die Studien wurden speziell für die Fragestellung dieser Arbeit konzipiert. Die wissenschaftliche Leitung des Projekts bildeten zuerst Rainer Peek und Wilfried Bos, ab Juli 2009 Isabell van Ackeren und Wilfried Bos. Finanziert wurde das Projekt durch ein Stipendium der Friedrich-Ebert-Stiftung sowie aus Mitteln der Arbeitsgruppe Bildungsforschung der Universität Duisburg-Essen und des Instituts für Schulentwicklungsforschung der Technischen Universität Dortmund. Das Ministerium für Schule und Weiterbildung in Nordrhein-Westfalen, welches dort für die Projekte VERA3 und Lernstand8/VERA8 zuständig ist, wurde vorab aber über die Befragung informiert, war aber nicht eingebunden.

Projektbeginn war im März 2009. Aufbauend auf einer im Jahr 2008 durchgeführten Vorstudie mit qualitativen Interviews (s. 3.4.2) wurde das Projekt problemorientiert konzipiert, d.h. der Ausgangspunkt war „Testcoaching“ als ein für sich stehender Begriff, der erst nach und nach einerseits als Unterrichtsqualität (Studie A) und andererseits als Form der Reaktion auf (angekündigtes) Feedback (Studie B) ausdifferenziert und dem Lehrer-Expertenansatz verbunden wurde. Oberstes Ziel der Studien war, das Vorbereitungsverhalten möglichst flächendeckend in Nordrhein-Westfalen abzubilden, sodass qualitative Ansätze ausschieden. Der Kontext Lernstand8 wurde gewählt, da zu VERA3 länger schon Untersuchungen zu den Auswirkungen auf die Unterrichtsqualität vorgenommen werden, wenngleich diese nicht speziell auf eine Unterrichtsqualität im Sinne von Testcoaching ausgelegt sind (Koch et al., 2006).

Die in den Kapiteln 2, 3 und 4 dargelegten theoretischen Konzeptionen und die bisherigen Befunde zu zentralen Vergleichsarbeiten, Testcoaching und Lehrerprofessionalität werden nun durch die Fragestellungen dieser Arbeit zusammengeführt.

[91] Zur Reflexion der Instrumente siehe Abschnitt 6.2.

5 Forschungsfragen und Hypothesen

Die Konzeption von Vergleichsarbeiten gleichzeitig als Instrument der Qualitätsentwicklung als auch der Rechenschaftslegung (s. Kap. 2) lässt in Verbindung mit den Befunden in anderen Staaten zur Wirkung von standardisierter Schulleistungsmessung vermuten, dass die Einführung von Vergleichsarbeiten flächendeckendes Testcoaching als Vorbereitung auf diese zentrale Schulleistungsmessung initiiert hat. Obwohl die Ergebnisse aus zentralen Vergleichsarbeiten nicht die gleichen Konsequenzen (z.B. geringere finanzielle Ressourcen für einzelne Schulen oder Schulschließungen) herbeiführen können wie High-Stake-Tests, kann unter Berücksichtigung der Ergebnisse aus Neuseeland und der im Vorfeld durchgeführten Interviewstudie ein sichtbares Ausmaß von entsprechenden Steuerungseffekten durch die Einführung von zentralen Vergleichsarbeiten angenommen werden. Da die Interviewstudie aber nur mit 17 Lehrkräften und nur im Raum Dortmund durchgeführt wurde, besteht Bedarf an umfangreicherer Aufklärung über das Ausmaß der Vorbereitung auf die zentralen Vergleichsarbeiten. Die erste zentrale Frage lautet folglich:

F1: In welchem Umfang und in welcher Qualität bereiten Lehrkräfte ihre Schülerinnen und Schüler auf die zentralen Vergleichsarbeiten vor*?*

Die Forschungsfrage lässt sich unter inhaltlichen Gesichtspunkten in weitere Forschungsfragen teilen. In Kapitel 3 wurde weitergehend dargestellt, dass sich Testcoaching nicht nur nach der Art des Inhalts klassifizieren lässt (Familiarity Approach, Content Approach und Test Wiseness Approach), sondern abhängig von der Qualität auch mit verschiedenen Zielen assoziieren lässt. Testcoaching kann je nach Qualität einer Täuschung über das korrekte Leistungspotenzial dienen (ein nicht-intendierter Effekt) oder aber die Testvalidität erhöhen (ein intendierter Effekt). Zusätzlich kann eine Vorbereitung auch Anlass für Übungsphasen wichtiger oder bisher nicht ausreichend von den Schülern verstandener Bereiche sein (ebenfalls ein intendierter Effekt). Die an F1 anschließenden Forschungsfragen lauten daher:

F1.1: Handelte es sich bei den in der Vorbereitung durchgeführten Maßnahmen eher um Maßnahmen, denen intendierte Effekte zugeschrieben werden können, oder eher um Maßnahmen, denen nicht-intendierte Effekte zugeschrieben werden?

F1.2: Inwieweit wird eine Testkompetenz vermittelt, die die Testvalidität erhöht?

F.1.3: Inwieweit wird die Vorbereitung genutzt, wichtige fachliche Inhalte zu vermitteln bzw. zu wiederholen?

Die bisherigen Forschungsfragen ergeben sich unabhängig davon, welches Steuerungsparadigma zugrunde gelegt wird. In Kapitel 2 wurde erläutert, dass zentrale Vergleichsarbeiten im Sinne der Neuen Steuerung auf den Unterricht wirken sollen. Unterstellt wird dabei eine output-orientierte Steuerung, bei der der Input und der Prozess von der Einzelschule bzw. den Lehrkräften mitgestaltet werden. Groß Ophoff vermutet aber, dass ein Großteil der in der Lehrerbefragungen zu zentralen Vergleichsarbeiten berichteten Unterrichtsentwicklungsmaßnahmen durch die Einführung selbst angestoßen wurden, aber keinem Reflexions- oder weitergehende Prozess zu Rückmeldeergebnissen entspringt (Groß Ophoff, 2013). Diemer und Kuper (2011) haben herausgearbeitet, dass das intendierte Reflexionsverfahren an Schulen nicht vollständig umgesetzt wird. Statt einer zweckprogrammierenden Nutzung der Ergebnisse aus zentralen Vergleichsarbeiten fanden sie vorwiegend konditionalprogrammierende Nutzungen[92]. Diese Fragestellung lässt sich auch auf die Vorbereitung übertragen. U.a. stellt für die Vorbereitung die Vorbereitungsintensität (zeitlicher Umfang und Anzahl der durchgeführten Maßnahmen) ein wichtiges Kriterium dar. Daraus ergibt sich die Frage:

F1.4a Inwieweit hat vorherige Erfahrung mit zentralen Vergleichsarbeiten in den Jahren davor einen Einfluss auf die Intensität der Vorbereitung?

Wenngleich Erfahrung zwar kein hinreichendes Kriterium für einen hohen Expertengrad ist, steigt doch mit zunehmender Erfahrung die Wahrscheinlichkeit für notwendige Lerngelegenheiten. Auch die Vorbereitung auf zentrale Prüfungen erfordert eine gewisse Kompetenz. Nicht jedem wird bewusst sein, dass unbekannte Aufgabenformate die Testvalidität reduzieren, andersherum kann eine Lehrkraft auch zu der Erkenntnis kommen, dass zu viel Unterrichtszeit für die Vorbereitung aufgewendet wurde oder die falschen Maßnahmen gewählt wurden. In diesem Zusammenhang können auch die vorherigen Forschungsfragen erneut betrachtet werden und Vergleiche zwischen Lehrkräften mit VERA8-Erfahrung und Lehrkräften ohne VERA8-Erfahrung durchgeführt werden.

[92] „Zweckprogrammierung“ meinte in dem Fall die reflexive Nutzung der Ergebnisse aus zentralen Vergleichsarbeiten. Unterrichtsentwicklung findet dann als Änderungen des Unterrichts nach den zentralen Vergleichsarbeiten statt, weil sich die Änderungen als Reflexionsergebnis darstellen. Von „Konditionalprogrammierung“ sprechen Diemer und Kuper (2011) hingen, wenn die Änderungen keiner Reflexion der rückgemeldeten Ergebnisse entspringen, sondern stattdessen aus antizipierten Testergebnissen.

In der vorab durchgeführten Interviewstudie äußerten die befragten Lehrkräfte zum Teil, einen Schwerpunkt auf die Teilkompetenzen gelegt zu haben, die schwerpunktmäßig in den Test von Lernstand8 im jeweiligen Schuljahr getestet werden sollten. Die Schwerpunktsetzung im Unterricht wurde an manchen Schulen auch innerhalb der Fachkonferenzen abgesprochen. Für das Fach Mathematik, auf welches sich die nachfolgenden Studien beziehen, gab es diese Schwerpunktsetzung bei den Tests zu VERA8 in den Jahren 2009 bis 2012 nicht mehr. Im Sinne einer Output-Orientierung können Lehrkräfte aber trotzdem einen besonderen Schwerpunkt bei einer inhaltsbezogenen oder prozessbezogenen Kompetenz[93] setzen, wenn sie bei dieser noch größeren Lernbedarf sehen. Interessant ist dann:

F 1.4b: Inwieweit werden Schwerpunktsetzungen mit Blick auf VERA8 vorgenommen?

Zur Beurteilung der Vorbereitung auf zentrale Vergleichsarbeiten spielt die Frage, ob die Maßnahmen zweckprogrammierenden Überlegungen entstammen, allerdings nicht allein eine Rolle. Im Zusammenhang mit der vorab durchgeführten Interviewstudie (Hahn, 2008) ist zu erwarten, dass für eine Vorbereitung auf die zentralen Vergleichsarbeiten entsprechendes begleitendes Material der gängigen Schulbuchverlage (Vorbereitungshefte und Kompetenzhefte/Lernhilfen) im Unterricht eingesetzt wird. Hierbei handelt es sich in erster Linie nicht um eine Frage nach Zweck- oder Konditionalprogrammierung, sondern um die Frage, inwiefern durch den Einsatz der Hefte im Unterricht doch erneut eine input-orientierte Steuerung vorliegt. Vorbereitungshefte zeichnen sich durch den starken Aufgabenbezug aus. Statt einer output-orientierten Steuerung liegt dann eine neue input-orientierte Steuerung vor, wenn die Einführung von zentralen Vergleichsarbeiten zu einer flächendeckenden Nutzung dieser Hefte geführt hat und Lehrkräfte der Ansicht sind, diese wegen der zentralen Vergleichsarbeiten notwendigerweise einsetzen zu müssen. Alternativ können Lehrkräfte statt auf aufgabenbezogene Vorbereitungshefte auch auf Lernhilfen/Kompetenzhefte zurückgreifen. Diese sind nicht auf die Bewältigung von zentralen Vergleichsarbeiten ausgelegt, sondern sollen bei Wiederholungsphasen unterstützen und fokussieren auf die Bildungsstandards und neuen (Kern-)Lernpläne. Wird hierauf zurückgegriffen, kann durchaus von einer output-orientierten Steuerung gesprochen werden, da die Nutzung von Lernhilfen nicht direkt nahe gelegt wird. Die Forschungsfrage dazu laute:

F1.5: Spricht der Umfang der Nutzung von Vorbereitungsheften und Kompetenzheften/Lernhilfen eher für eine output- oder eine input-orientierte Steuerung?

[93] Die Kernlehrpläne für das Fach Mathematik unterscheiden vier inhaltsbezogene Kompetenzen (Arithmetik/Algebra, Funktionen, Geometrie und Stochastik) und vier prozessbezogene Kompetenzen (Argumentieren/Kommunizieren, Problemlösen, Modellieren, Werkzeuge)

Zentrale Vergleichsarbeiten und Testcoaching sind damit in einem ersten Schritt verwoben. Die nachfolgenden Forschungsfragen und Hypothesen versuchen, das berichtete Vorbereitungsverhalten durch differentielle Modelle zu erklären. Dadurch wird u.a. beabsichtigt, die Frage zu beantworten, ob das Vorbereitungsverhalten eher konditional- oder zweckprogrammierenden Überlegungen zuzurechnen ist. Zuerst werden Typenunterschiede im Sinne von konditionalprogrammierendem Verhalten betrachtet. Die Vergleichsarbeiten werden als Ausgangspunkt für vorher durchgeführte Vorbereitung verstanden.

In Kapitel 3 und Kapitel 4 wurde deutlich, dass Testcoaching als im Unterricht durchgeführte Form der Vorbereitung auf Schulleistungsmessungen eine bestimmte Form der Unterrichtsgestaltung darstellen und Unterrichtsqualität im Rahmen des Lehrer-Expertenansatzes analysiert werden kann. Dem liegt die Vorstellung einer Anforderung-Ressourcen-Verbindung zugrunde, wie sie für berufliche Situationen allgemein im Job-Demand-Resource-Model von Schaufeli und Kollegen abgebildet wird. Schüler und Schülerinnen auf Vergleichsarbeiten vorzubereiten, stellt als eine bestimmte Form der Unterrichtsqualität eine konkrete Anforderung dar, denen sich Lehrkräfte gegenübergestellt sehen können. Der Grad der Anforderungserfüllung hängt nach dem JD-R-Model von den persönlich zur Verfügung stehenden Ressourcen ab. Als globale Ressource für die Bewältigung der Unterrichtsanforderungen kann die Lehrer-Handlungskompetenz gesehen werden, die von Baumert und Kunter in ihrem Modell der Lehrer-Handlungskompetenz beschrieben und im COACTIV-Projekt (beispielsweise von Klusmann u.a.) als Erklärungsgröße untersucht wurde. Ähnlich dazu verhält es sich mit dem erweiterten Modell, welches am Ende von Kapitel 4 vorgestellt wurde und weitere Befunde zu diesem Bereich integriert. Eng zum ursprünglichen JD-R-Model kommt den personenbezogenen Überzeugungen eine große Bedeutung als notwendige Ressourcengrundlage für die Unterrichtsgestaltung zu. Untersuchungen von Schaarschmidt und Kollegen legen dabei eine Typisierung nach personenbezogenen Überzeugungen im beruflichen Kontext nahe, die von Klusmann u.a. erfolgreich empirisch untermauert werden konnte. Aber auch die Bedeutung der Kompetenz- und Kontrollüberzeugungen konnte mehrfach untermauert werden. Daraus resultiert die nächste Forschungsfrage:

F2: Inwieweit lassen sich Unterschiede zwischen Lehrkräften in Umfang und Qualität der Vorbereitung durch die Typisierung der Lehrkräfte nach personenbezogenen Überzeugungen erklären?

Dem voraus geht eine weitere Forschungsfrage:

F2.1: Inwieweit lassen sich die Lehrkräfte ähnlich klassifizieren wie in den Befunde von Schaarschmit u.a. bzw. Klusmann u.a. für die personenbezogenen Überzeugungen im beruflichen Kontext, wenn zusätzlich Kompetenz- und Kontrollüberzeugungen berücksichtigt werden?

Verknüpft sind damit die Hypothesen, die einen hohen Ressourceneinsatz mit einer intensiven Vorbereitung verknüpfen und die Vorbereitung dabei qualitativ und bzgl. des Umfangs unterscheiden.

H2.1: Lehrkräfte, die überdurchschnittlich engagiert sind (Typ G[94] oder Typ A), bereiten umfangreicher auf die Vergleichsarbeiten vor als Lehrkräfte, die über weniger Arbeitsengagement verfügen (Typ B und Typ S).

Eine umfangreichere Vorbereitung kann sich folglich in einem größeren Stundenvolumen oder aber auch in mehr behandelten Themen ausdrücken. Dementsprechend kann dies noch einmal für den Typ G spezieller formuliert werden:

H2.2: Lehrkräfte, die über umfangreichere Ressourcen im Sinne des Modells der Lehrer-Handlungskompetenz nach Baumert und Kunter verfügen (ähnlich dem Typ G), bereiten qualitativ besser auf die Vergleichsarbeiten vor als andere Lehrkräfte.

Qualitativ besser meint vier verschiedene Elemente: (a) Die Vorbereitung ist anspruchsvoller. (b) Die Vorbereitung umfasst neben den inhaltsbezogenen Kompetenzen auch prozessbezogene Kompetenzen. (c) Die Vorbereitung lässt Schülerinnen und Schülern Freiraum in der Vorbereitung und unterstützt sie dabei. (d) Die Vorbereitung ist variantenreicher. Diese vier Qualitätsmerkmale sind parallel zu Kriterien guten Unterrichts formuliert wie man sie in einschlägigen Listen von Brophy (1999), Helmke (2009) oder Meyer (2009) findet und entsprechen der Forderung nach herausfordernden Lerngelegenheiten, Orientierung am Lehrplan, Übung und Anwendung, Lehren von Strategien (Brophy, 1999) sowie Methodenvarianz (Meyer, 2009). Angenommen wird eine herausragende Stellung von Lehrkräften ähnlich des G-Typs, denkbar ist aber auch, dass ebenfalls Lehrkräfte des A-Typs

[94] Zur einfachen Darstellung werden analoge Bezeichnungen zur Klassifikation von Schaarschmidt und Fischer verwendet, auch wenn diese inhaltlich anders zusammengesetzt sind.

eine qualitativ besser Vorbereitung leisten, da diese sich von G-Typ-Lehrkräften vorwiegend durch die negativen beruflichen Emotionen unterscheiden.

Aus dem unterstellten Zusammenhang von ausreichenden Ressourcen und qualitativ hoher Vorbereitungsphase lassen sich umgekehrt auch Hypothesen über Lehrkräfte mit weniger benötigten Ressourcen folgern:

H2.3: Lehrkräfte, die über negative Kompetenz- und Kontrollüberzeugungen verfügen, bereiten auf kurzfristige Erfolge ausgerichtet auf die Vergleichsarbeiten vor.

Bei H2.3 wird unterstellt, dass Lehrkräfte mit negativen Kompetenz- und Kontrollüberzeugungen nicht erwarten, ihre Schülerinnen und Schüler durch ihren normalen Unterricht ausreichend vorzubereiten. Ihnen bleiben trotzdem zwei Optionen, zu passablen Ergebnissen zu gelangen, nämlich durch Auslagerung der Vorbereitung in die Verantwortung der Schüler und durch Tipps & Tricks, die vor allem auf gute Ergebnisse abzielen.

Prinzipiell kann eine zeitlich umfangreiche Vorbereitung auch durch einen geringen Einsatz von Ressourcen durchgeführt werden, wenn dabei vermehrt auf Vorbereitungshefte und außerunterrichtliche Vorbereitung zurückgegriffen wird.

H2.4: Lehrkräfte, die unterdurchschnittlich wenig Ressourcen einsetzen wollen (Typ S), verlagern die Vorbereitung in den Verantwortungsbereich der Schüler und ***bereiten vorwiegend nur mit Vorbereitungsheften auf die Vergleichsarbeiten vor.***

Da eine Vorbereitung auf die Vergleichsarbeiten aufgrund der testtheoretischen Anlage (vgl. Kap. 2) eher geringe Effekte haben wird bzw. diese schlecht zu beobachten sind, sollten Lehrkräfte außerhalb des optimalen Typs verstärkter in ihren Mustern verharren. Dies bedeutet konkret:

H2.5: Lehrkräfte, die unterdurchschnittlich wenige Ressourcen einsetzen wollen oder können (Typ B oder Typ S), bereiten weniger umfangreich auf die Vergleichsarbeiten vor als in den Jahren zuvor.

Auch die den Vergleichsarbeiten beigemessene Bedeutung sollte vom Ressourceneinsatz abhängen. Da die Vergleichsarbeiten als Steuerungsinstrument in frühere Gewohnheiten und alltägliche Abläufe eingreifen, können folgende Zusammenhänge vermutet werden:

H2.6: Lehrkräfte, die unterdurchschnittlich wenige Ressourcen einsetzen wollen oder können (Typ B oder Typ S), beschreiben die Vergleichsarbeiten als relativ unbedeutend.

Mit diesen sechs Hypothesen werden mögliche Zusammenhänge zwischen Testvorbereitung und Unterrichtsqualität abgedeckt, sofern es sich um Zusammenhänge handelt, die auf die Typisierung nach personenbezogenen Ressourcen zurückgehen.

Während die Forschungsfragen F2 und F2.1 die Vergleichsarbeiten lediglich als Anlass für Testcoaching nutzen, thematisiert der zweite in dieser Arbeit aufgegriffene Forschungsansatz wiederum stärker die Bedeutung der Vergleichsarbeiten als Instrument der Qualitätsentwicklung und Rechenschaftslegung. Die an Lehrkräfte im Rahmen der Vergleichsarbeiten gegebenen Rückmeldungen können als Feedback aufgefasst werden. Zentral ist hier das Modell zur pädagogischen Nutzung der Ergebnisse aus Vergleichsarbeiten von Helmke und Hosenfeld zu nennen, dass eine idealtypische Nutzung von Feedbackinformationen beschreiben soll.

Dem Modell nach setzt Unterrichtsentwicklung die Reflexion und diese wiederum die Rezeption von Leistungsrückmeldungen voraus. Diese als für die Rezeptionsforschung charakterisierende Herangehensweise betrachtet Testcoaching im Rahmen des zyklischen Modells als vorhergehende Pro-Aktion oder resultierende Re-Aktion. Testcoaching als vorhergehende Pro-Aktion ist hier ein Mittel, um negativem Feedback zu entgehen oder aber der Feedbackbotschaft zu mehr Validität zu verhelfen. Scheinbar muss zwischen den persönlichen Zielen im Rahmen der Qualitätsentwicklung und der Rechenschaftslegung abgewogen werden. Je mehr auf die Vergleichsarbeiten vorbereitet wird, desto unwahrscheinlicher sind schlechte Ergebnisse, aber auch desto mehr wird das Testergebnis verfälscht. Als Re-Aktion ist die Testvorbereitung nur ein Teil des normalen Unterrichts und möglicherweise das Ergebnis eines Reflexionsprozesses bzw. ein Resultat aus vorherigen VERA8-Ergebnissen der Lehrkraft (oder ggf. anderer Lehrkräfte aus dem Umfeld der Lehrkraft).

Hier offenbar sich zwei Lücken im Modell von Helmke und Hosenfeld. Die erste Lücke ergibt sich aus der Tatsache, dass das Modell mit seinen komplexen Einflüssen in Wirklichkeit kein zyklisches Modell ist (vgl. auch Maier, 2009), sondern mit der Rezeption beginnt. Unterricht vor der Evaluation findet keine Berücksichtigung, folglich auch keine Testvorbereitung als Pro-Aktion. Die zweite Lücke stellt die fehlende „Exit-Möglichkeit“ im Anschluss an die Rezeption dar. Wenn die zentralen Vergleichsarbeiten gute Ergebnisse hervorbringen, besteht kein Handlungsbedarf, sodass dann weder eine Ursachenreflexion noch eine

anschließende Re-Aktion sinnvoll scheint. Statt „pädagogische Nutzung von Vergleichsarbeiten“ beschreibt das Modell nur den Prozess der Unterrichtsentwicklung nach nicht ausreichenden Ergebnissen in zentralen Vergleichsarbeiten. Entsprechend deutet auch Hosenfeld (2010) die Befunde ihrer latenten Klassenanalysen.

Befunde aus der allgemeinen Feedbackforschung, die in den Theorien von Kluger und DeNisi bzw. Hattie und Timperley abgebildet werden, und Befunde aus der Rezeptionsforschung zu Vergleichsarbeiten von Hosenfeld (ebenda) regen eine Typisierung an, die die aufgabenbezogene Selbstwirksamkeitserwartung, die Attributionsüberzeugungen und die Klarheit und Internalisierung der Ziele[95] sowie die Nützlichkeitseinschätzung und die Rezeptions- und Reflexionsbereitschaft im Kontext der Vergleichsarbeiten betrachten. Auch die durch die Schulleitung und Kollegen angebotene Unterstützung sollte einen Einfluss haben. Parallel zum ersten Zugang lauten die Forschungsfragen F3, F3.1 und F3.2:

F3: Inwieweit lassen sich Unterschiede zwischen Lehrkräften in Umfang und Qualität der Vorbereitung durch die Typisierung der Lehrkräfte nach Einstellungen zu und Umgang mit Rückmeldungen aus den Vergleichsarbeiten erklären?

und

F3.1: Inwieweit lassen sich die Lehrkräfte sinnvoll über die unterrichtsbezogene Selbstwirksamkeitserwartung, die Gewissenhaftigkeit, die Attributionsüberzeugungen und Internalisierung der Unterrichtsziele sowie erlebte Unterstützung klassifizieren*?***

und

F3.2: Inwieweit lassen sich die Lehrkräfte sinnvoll über die Nutzung von Daten aus Vergleichsarbeiten aus vorherigen Jahren klassifizieren?

Der allgemeinen Feedbackforschung nach sollten Lehrkräfte mit hoher Selbstwirksamkeitserwartung und großer erlebter Unterstützung sich prinzipiell eher

[95] Die Klarheit und Internalisierung der Ziele können als dritte Lücke im Modell von Helmke und Hosenfeld identifiziert werden. Die Klarheit der Ziele ist u.a. Voraussetzung für die Rezeption der Ergebnisse, ohne diese können die Ergebnisse nicht verstanden werden. Die Internalisierung muss wiederum gegeben sein, wenn Reflexions- und Änderungsprozesse angestoßen werden sollen.

Feedbacksituationen aussetzen, weil sie sich positives Feedback versprechen oder annehmen, die nötigen Schritte zur Korrektur vornehmen zu können. Die Voraussetzung, um überhaupt Rückmeldungen aus Vergleichsarbeiten als Feedback über den Unterrichtserfolg anzusehen, ist allerdings die Schülerleistungen in der Verantwortung der Lehrkräfte zu erkennen (Attributionsüberzeugungen) und die Kernlehrpläne bzw. Bildungsstandards (auf denen die Vergleichsarbeiten aufbauen) als Unterrichtsziele zu akzeptieren (Klarheit und Internalisierung von Zielen). Demnach ergeben sich für die Forschungsfrage F3.1 vier sinnvoll interpretierbare Muster: Typ F21.A Lehrkräfte, die eine hohe SWE besitzen, große Unterstützung wahrnehmen und sowohl die Schülerleistungen in ihrer Verantwortung verorten als auch die Unterrichtsvorgaben akzeptieren, Typ F21.B Lehrkräfte, die sowohl die Schülerleistungen in ihrer Verantwortung verorten als auch die Unterrichtsvorgaben akzeptieren, aber über eine geringe SWE oder erlebte Unterstützung verfügen, Typ F21.C Lehrkräfte, die zumindest entweder die Schülerleistungen nicht in ihrer Verantwortung verorten oder auch die Unterrichtsvorgaben nicht akzeptieren und über eine geringe SWE und erlebte Unterstützung verfügen und Typ F21.D Lehrkräfte, die zwar eine hohe SWE und erlebte Unterstützung besitzen, aber die Vergleichsarbeiten nicht als Rückmeldung über den Unterrichtserfolg ansehen oder/und die Bildungsstandards nicht akzeptieren.

Bezogen auf einen möglichen Feedbackprozess bedeutet die Vorbereitung auf zentrale Vergleichsarbeiten, dass eine geringe Vorbereitung im Sinne eines Familiarity Approach sinnvoll ist. Eine intensivere Vorbereitung aber führt zu einem überhöhten Testscore und lässt die rückgemeldeten Ergebnisse aus VERA8 als Feedback für die Lehrkraft unbrauchbar werden. Es wird folglich angenommen:

H3.1: Lehrkräfte, die Schülerleistungen nicht durch sich selbst verantwortet sehen oder/und die Bildungsstandards bzw. Kernlehrpläne als Ziele nicht internalisiert haben (Typ F21.C, Typ F21.D), bereiten weniger umfangreich auf die Vergleichsarbeiten vor.

Bei der Hypothese H3.1 wird unterstellt, dass Lehrkräfte generell eher dazu neigen, auf zentrale Vergleichsarbeiten vorzubereiten, diese ein Feedback produzieren, welches generell das Selbstbild gefährden könnte. Wenn man die Schülerleistungen aber nicht in seiner Verantwortung sieht oder die Ziele, die den Tests zugrunde liegen, nicht internalisiert hat, existieren andere Möglichkeiten als eine intensive Testvorbereitung, um den Angriff auf sein Selbstbild abzuwenden.

H3.2: Lehrkräfte, die sowohl die Schülerleistungen in ihrer Verantwortung verorten als auch die Unterrichtsvorgaben akzeptieren und eine hohe SWE besitzen (Typ F21.A), bereiten minimal vor.

H3.3: Lehrkräfte, die sowohl die Schülerleistungen in ihrer Verantwortung verorten als auch die Unterrichtsvorgaben akzeptieren, aber eine niedrige SWE besitzen (Typ F21.B), bereiten am umfangreichsten vor.

Mit der Hypothese H3.4 soll das Verhalten genau derjenigen Lehrkräfte beschrieben werden, die sich durch zentrale Vergleichsarbeiten dazu genötigt sehen, die zu erwartende Rückmeldung positiver erscheinen zu lassen als es dem tatsächlichen Leistungsstand ihrer Schüler entspricht.

Neben den Elementen, die die FIT von Kluger und DeNisi als für die Feedbacknutzung als relevant darstellen und mit denen nicht unterschieden wird, ob die evtl. Vorbereitung ein Feed-Forward-Effekt oder ein Effekt eines Reflexionsprozesses vergangener zentraler Vergleichsarbeiten ist, wird mit dem Modell von Helmke und Hosenfeld konkret ein Rezeptions- und Reflexionsprozess unterstellt. In Anlehnung an die von Hosenfeld gefundene Klassifikation scheinen drei Typen sinnvoll begründet:

Typ R22.A zeigt eine hohe Rezeption und Evaluationsbereitschaft und (in Folge von unerwartet schlechten Ergebnissen) auch eine hohe Reflexionsintensität und Veränderungsbereitschaft. Typ R22.B zeigt ebenfalls eine hohe Rezeption und Evaluationsbereitschaft, aufgrund von erwarteten Ergebnissen entfällt aber der Grund für eine Reflexionsphase. Dieser Typ ist also abweichend vom Modell von Helmke und Hosenfeld, nachdem eine hohe Nutzung der Rückmeldungen sich stets in einem umfangreichen Reflexionsprozess ausdrückt. Typ R22.C besitzt hingegen keine große Evaluationsbereitschaft, rezipiert die Ergebnisse nur eine eingeschränkt und berichtet folgend auch nur von einer eingeschränkten Reflexionsphase.

Es wird allerdings weiter angenommen, dass die drei Reflexionstypen nicht für sich stehen, sondern die Nutzung von Daten aus zentralen Vergleichsarbeiten ein Resultat dessen sind, wie stark die Bedingungen erfüllt sind, um sich einem Feedbackprozess auszusetzen. Der Typ R22.A hat dieselben Bedingungen als Voraussetzung wie der Typ F21.A, möglicherweise kommen auch Personen in Frage, die als F21.B klassifiziert wurden, wenn zwar die eigene Ressourcen gering sind, aber Unterstützung durch die Schulleitung oder andere vorhanden ist. Der Typ R22.B hat ebenfalls große eigene Ressourcen zur Bedingung, denn nur mit diesen ist eine entsprechende Unterrichtsqualität unabhängig von sehr guten äußeren Bedingungen möglich, sodass auch hier mehrheitlich der Typ F21.A als Klassifikationsgruppe angenommen werden kann. In der Klasse Typ R22.C sollten folglich diejenigen zu finden sein, die sich durch eine geringe Akzeptanz der Ziele auszeichnen F21.D und evtl. zusätzlich auch wenige förderliche Ressourcen besitzen F21.C. Damit ergeben sich statt der ursprünglichen drei Reflexionstypen möglicherweise vier, da der Typ R22.C in zwei unterschiedliche Feedbacktypen zerfallen sollte.

Die zugehörigen Hypothesen lauten:

H3.4: Lehrkräfte, die hohe Werte bei der Rezeptions-, Reflexions- UND Veränderungsbereitschaft im Kontext der Vergleichsarbeiten zeigen, bereiten intensiv vor.

Lehrkräfte dieses Typs sollten eine intensive Vorbereitung durchführen, weil sich einerseits dem durch VERA8 initiierten Feedbackprozess stellen, andererseits aber negative Resultate befürchten müssen. Welche Maßnahmen sie dabei vor allem nutzen, kann nicht vorhergesagt werden. Möglich ist sowohl, dass die vorher unerwartete schlechten Ergebnisse auf Schwierigkeiten mit den Aufgabenformaten zurückgeführt werden und die Vorbereitung den Schwerpunkt auf ein Familiarity Approach legt, als auch, dass inhaltliche Lücken als Ursache entdeckt wurden und vermehrt Wiederholungsphasen vorgesehen werden.

H3.5: Lehrkräfte, die dem Typ R.22.B und Typ R.22.C zugeordnet werden, sollten minimal vorbereiten.

Die anderen Typen hingegen sollten nicht oder nur minimal vorbereiten, weil sie auch ohne eine intensive Vorbereitung positive Ergebnisse erwarten R22.B oder VERA8 nicht als Feedbackprozess annehmen R22.C.

6 Methodologie

Im Folgenden wird die Anlage der beiden dieser Dissertation zugrundeliegenden Studien diskutiert. Der erste Abschnitt diskutiert die Wahl der Datengrundlage mit der Begründung für die Wahl der Schulform und des Unterrichtsfachs und beinhaltet sowohl eine Gegenüberstellung von den gewählten schriftlichen Befragungen im Vergleich zu Unterrichtsbeobachtungen und Leistungstests als auch eine Gegenüberstellung von einerseits postalischen und andererseits rein webbasierten schriftlichen Befragungen. Außerdem werden eine Beschreibung des versendeten Fragebogenpakets gegeben und die angestrebte und die realisierte Stichprobe dokumentiert. Im zweiten Abschnitt werden die beiden eingesetzten Fragebögen so konkret wie nötig dargestellt und ihr Aufbau dokumentiert. Dabei wird unterschieden zwischen den Fragebogenitems zum Vorbereitungsverhalten, die in beiden Studien identisch waren, den Items zum erweiterten Modell der Lehrer-Handlungskompetenz (Studie A) und den Items zu VERA8 als Feedbackinstrument (Studie A & B). Der letzte Abschnitt bietet einen Einblick in das Analysevorgehen der gewonnenen Datensätze. Dies umfasst sowohl den deskriptiven Teil als auch die Skalenbildung und Vergleiche der Experten-Modelle mittels latenter Klassenanalysen.

6.1 Design der Studien

6.1.1 Wahl der Schulform und des Unterrichtsfachs

Die Studien wurden zur Reduzierung der Komplexität auf Gymnasien beschränkt. Haupt- und Realschulen waren zum Erhebungszeitpunkt ihrer Größe nach meistens derart angelegt, dass dieselben Lehrkräfte sowohl in der achten als auch in der zehnten Jahrgangsstufe unterrichteten und dadurch zusätzlich durch die Vorbereitung auf die Zentralen Abschlussprüfungen 10 eingespannt waren. Um die Wahrscheinlichkeit zu reduzieren, dass Lehrkräfte beide Vorbereitungsphasen verwechselten, während sie rückwirkend befragt wurden, wurde auf andere Schulformen zurückgegriffen. Die Entscheidung fiel für Gymnasien und gegen Gesamtschulen wegen der anderenfalls zusätzlich zu berücksichtigenden innerschulischen Differenzierung zwischen Grund- und Erweiterungskursen. Mit der Einführung von eigenen Testheften für Gymnasien sind Schulformeffekte abgemindert, sofern sie aus Bodeneffekten der Testhefte resultieren. Schulformeffekte aufgrund unterschiedlicher Ausbildungswege oder aufgrund des Habitus

der Gymnasiallehrkräfte bleiben hingegen bestehen. Dies kann sich sowohl auf die Unterrichtsgestaltung als auch auf den Umgang mit Veränderungsprozessen auswirken. Die Ergebnisse sind daher nur eingeschränkt auf andere Schulformen und auf die Wirkung von VERA3 auf Grundschullehrkräfte und den Unterricht in Grundschulen übertragbar. Grundsätzlich sind die theoretischen Basiskonzepte aber schulformübergreifend angelegt (vgl. hierzu auch Abs. 8.2).

Die Einschränkung auf das Fach Mathematik ist den massiven Fehldrucken in den Testheften für Deutsch und Englisch im Jahr 2009 geschuldet, die einen Imageschaden des Instruments zentrale Vergleichsarbeiten befürchten ließen. Mögliche Verzerrungen durch die spezielle Situation im Jahr 2010 nach dem möglichen Vertrauensverlust in das Instrument sollten vermieden werden. Zusätzlich war durch die Einschränkung gewährleistet, dass alle Untersuchungsteilnehmenden zumindest in groben Zügen die Rückmeldeergebnisse aus VERA8 verstehen, da die nötigen Statistikkenntnisse Grundlage für den eigenen Unterricht sind. Dies ist (vgl. Kap 4) eine wichtige Voraussetzung für eine angemessene Nutzung der Ergebnisrückmeldungen zur Unterrichtsentwicklung.

Den beiden Studien dieses Projekts liegen schriftliche Befragungen zugrunde. Um das Projektziel zu erreichen, die Vorbereitung der Schülerinnen und Schüler auf VERA8 durch die Lehrkräfte zu erfassen und aus der Perspektive des Lehrer-Expertenansatzes zu erklären, wären verschiedene methodische Ansätze möglich gewesen. Neben schriftlichen Befragungen hätten auch Unterrichtsbeobachtungen und Leistungstests eingesetzt werden können. Der nächste Abschnitt soll daher die Abwägungen für die durchgeführte Methode, die schriftliche Befragung in postalischer wie webbasierter Form nachvollziehbar werden lassen.

6.1.2 Schriftliche Befragung vs. Unterrichtsbeobachtung und Leistungstest

Die Angaben der Lehrkräfte über das Vorbereitungsverhalten sind der zentrale Baustein in beiden Studien. Das Unterrichtsverhalten von Lehrkräften zu erfassen gilt grundsätzlich als schwierig. Beobachter-, Lehrer- und Schülerperspektive gelangen häufig zu schwach übereinstimmenden Befunden, wenn Unterrichtsqualität (beispielsweise als vorherrschende Disziplin, Bezugsnormorientierung oder Lehrstil) anhand von weichen Kriterien gemessen wird. Problematisch sind dabei u.a. fehlende Kriterien zur Bewertung insbesondere bei hoch-inferenten Unterrichtsmerkmalen und sozial erwünschte Antworttendenzen bei Lehrkräften (Clausen, 2002).

Die aus der Interviewstudie explizierten Qualitätsmerkmale des Vorbereitungsverhaltens werden von diesen Schwierigkeiten allerdings nicht tangiert. Erstens handelt es sich um

niedrig-inferente Merkmale. Diese können auch von Lehrkräften zuverlässig berichtet werden (Rakoczy, 2007). Abgefragt wurden die aufgewendeten Unterrichtsstunden, Vorbereitungsinhalte, Vorbereitungsmethoden und der Einsatz von Vorbereitungsheften und Lernhilfen. Die hohen Übereinstimmungen zwischen Lehrer- und Schülerberichten zum Vorbereitungsverhalten in der Interviewstudie stützt die allgemeinen Annahmen für die Zuverlässigkeit der Lehrerurteile in diesem konkreten Fall zusätzlich. Zweitens ist anzunehmen, dass der Befragungsgegenstand für die Befragten wertneutral ist. Eine Differenzierung in wünschenswerte und unerwünschte Qualitätsmerkmale der Vorbereitung auf die Vergleichsarbeiten setzt (durch jahrelange Erfahrung gewonnenes) Expertenwissen über Testcoaching voraus, welches für deutsche Lehrkräfte nicht angenommen werden muss. Zusätzlich sind positiv zu bewertende und negativ zu bewertende Vorbereitungsmerkmale im Fragebogen durchmischt. Drittens ist der tatsächliche Unterrichtsgegenstand dieses Projekts die geplante Unterrichtsqualität, da nicht die Wirkung der Vorbereitungsqualität auf die Schülerleistung, sondern die Wirkung der Vergleichsarbeiten auf die Vorbereitung untersucht wird. Aus diesem Blickwinkel macht es keinen bedeutenden Unterschied, ob die geplante Vorbereitung auch tatsächlich vollständig realisiert werden konnte. Als Antwortmöglichkeiten wurden dichotome oder dreistufige Alternativen angeboten. Dies erleichtert die Beantwortung zusätzlich.

Prinzipiell schwieriger, da hoch-inferent, sind Einschätzungen über die mögliche Veränderung der Vorbereitungsintensität und des wahrgenommenen Drucks, ausreichend auf die Vergleichsarbeiten vorzubereiten bzw. gute Ergebnisse zu erzielen. Hier fehlt es an klaren Kriterien, um die Vorbereitungsintensität und den erlebten Druck (insbesondere den in der Vergangenheit) angemessen messen und erinnern zu können. Gleiches gilt für Einschätzungen über die Bedeutung der Vergleichsarbeiten für die eigene Person, aber besonders die Einschätzung der Bedeutung für Schülerinnen und Schüler, Kollegen und Kolleginnen und die Schulleitung. Zusätzlich erschwerend wirkt hier, dass eine nicht zwingend zugängliche Innensicht der Schülerinnen und Schüler bzw. Kollegen und Kolleginnen erforderlich ist. Andererseits dienen die Fragen nach der wahrgenommenen Bedeutung der Vergleichsarbeiten im Umfeld auch zur Verifikation der Aussagen über den wahrgenommenen Druck, der mitunter durch diese aufgebaut wird. In den Fragebögen finden dadurch Rückversicherungen statt.

Als niedrig-inferent können auch die Fragen des zweiten Teils in Studie B zum Rezeptions-, Reflexions- und Veränderungshandeln betrachtet werden. Gleiches gilt für die Fragen zum Umgang der Schulleitung mit den Rückmeldeergebnissen, zum Medieninteresse nach den Rückmeldeergebnissen*[96] und zum Standorttyp der Schule*.

Das der Studie A zugrundeliegende Modell, das Modell der Lehrer-Handlungskompetenz nach Baumert und Kunter, umfasst einen immanenten Bereich „Wissen und Können“. Da Wissen und Können durch Selbstauskünfte nicht angemessen erfasst werden können, hätte es zusätzliche Erhebungen bedurft, bei denen zumindest Papier & Bleistift-Leistungstests

[96] * Dieser Teil der Daten wurde nicht in die Auswertung dieser Arbeit übernommen.

hätten eingesetzt werden müssen (Kunter & Klusmann, 2010). Neuweg stellt sogar auch diese in Frage und plädiert stattdessen dafür, die Fähigkeiten und Fertigkeiten von Lehrkräften auch in realen Situationen zu messen (Neuweg, 2008). Dazu wären Unterrichtsbeobachtungen oder Videographie nötig gewesen. In Anbetracht des vordergründigen Projektziels, das Vorbereitungsverhalten auf die Vergleichsarbeiten möglichst flächendeckend zu erfassen, waren solche Testverfahren zu aufwendig. Vor allem aber sind solche Verfahren nicht anonymisiert durchführbar, welches aber eine wichtige Voraussetzung bei der Erfassung von sensiblen Untersuchungsgegenständen wie dem Vorbereitungsverhalten ist. Ein realitätsnäheres, etwas weniger aufwendiges und weniger sensibles Verfahren können Tests mithilfe von Filmvignetten sein (Oser, Heinzer & Salzmann, 2010), aber auch diese Verfahren sind nur onlinegestützt anonym. Evtl. wäre dadurch ein Teil der Lehrkräfte systematisch ausgeschlossen gewesen (vgl. Coverage-Effekte), weil es an den technischen Möglichkeiten fehlte, sich an onlinegestützten Testverfahren zu beteiligen, oder ihnen das zu aufwendig schien[97]. Eine Verbindung von Ergebnissen aus Leistungstests und Vorbereitungsverhalten war dadurch nicht möglich.

Bei den anderen drei Bereichen des Modells zur Lehrer-Handlungskompetenz handelt es sich um explizite Persönlichkeitseigenschaften oder verhaltensfernen Kognitionen. Auch in Studie B wurden zusätzlich explizite Persönlichkeitseigenschaften erfasst. Diese können (im Gegensatz beispielsweise zu impliziten Motivstrukturen) mittels Fragebogentechniken erfasst werden (Leuchter et al., 2006). Dazu wird in der Regel auf Beurteilungsaufgaben zurückgegriffen (Jankisz & Moosbrugger, 2008). Für beide Studien wurden Beurteilungsaufgaben mit sechsstufigen, verbalen, bipolaren oder unipolaren Rating-Skalen gewählt. Obwohl diese Skalen theoretisch nur Ordinalskalenniveau erreichen, hat es sich durchgesetzt den Untersuchungsteilnehmenden zu unterstellen, sie nutzten diese beim Ausfüllen wie Intervallskalen (Porst, 2009). Um diesen Effekt zu unterstützen, wurden nur die Endpunkte bezeichnet. Auf eine mittlere Antwortkategorie wurde verzichtet, um das Ausweichverhalten der Befragungsteilnehmenden gering zu halten. Mittelkategorien werden häufig missverstanden und Abweichungen davon als Normabweichung interpretiert, sodass die Mittelkategorie als sozial erwünscht betrachtet wird (Jankisz & Moosbrugger, 2008).

6.1.3 Zum Verhältnis von postalischer Befragung und Onlinebefragung in den Studien

Die Teilnahme an den schriftlichen Befragungen der Studien A und B war über zwei Wege möglich: Neben der Option, den ausgedruckten Fragebogen auszufüllen und mittels des beigelegten Antwortumschlags zurückzusenden (postalische Befragung), war im Anschreiben auch ein Link angegeben, unter dem der jeweilige Fragebogen online aufgerufen werden konnte (Webbefragung). Leeuw, Hox und Dillman (2008) empfehlen die Kombination von

[97] Online durchgeführte Leistungstests mit Filmvignetten erfordern umfassendere technische Voraussetzungen als HTML-basierte Befragungen.

verschiedenen Methoden der schriftlichen Befragung, um Stichprobenfehler zu reduzieren, warnen aber davor, dass die Kombination zweier Methoden Messfehler (Methodenfehler) die Datengüte mindern können. Durch das parallele Angebot der beiden Befragungsmethoden sollte in erster Linie die Rücklaufquote erhöht werden, da über den Link zur Webbefragung eine Teilnahme auch dann noch möglich war, wenn das ausgedruckte Exemplar des Fragebogens nicht mehr vorlag oder wenn nicht ausreichend viele Fragebögen an die Schule versendet worden waren. Außerdem konnten gleichzeitig die Zahl der einzulesenden Fragebögen und die Portokosten reduziert werden. Als Anreiz für die Teilnahme an der Webbefragung wurde zugesagt, für jeden online ausgefüllten Fragebogen 0,50€ an die Organisation „SOS Kinderdörfer“ zu spenden. Andersherum sollten durch die postalische Befragung auch diejenigen Lehrkräfte angesprochen werden, die Vorbehalte gegen Webbefragungen haben oder für die die Teilnahme an postalischen Befragungen besser möglich ist. Der Vorteil von Webbefragungen liegt vor allem in den umfangreichen Gestaltungsmöglichkeiten. Diese lassen im Vergleich zu postalischen Befragungen multimediale Darbietungen von Fragen zu, können aber auch relativ ähnlich zu postalischen Befragungen gestaltet sein. Webbefragungen weisen außerdem nur Fixkosten für die Erstellung der Befragung sowie die Hostingkosten auf. Schließlich wird anders als bei postalischen Befragungen der Umgang mit dem Fragebogen dokumentiert, sodass die Bearbeitungszeit der Untersuchungsteilnehmenden sowie evtl. Abbruchseiten eingesehen und für evtl. Korrekturen ausgewertet werden können. Webbefragungen weisen allerdings im Vergleich zu postalischen Befragungen einige besondere Herausforderungen auf und kombinierte schriftliche Befragungen können zu unterschiedlichen Antwortverteilungen führen, sodass stets geprüft werden muss, ob sie für die Untersuchung geeignet sind.

Es müssen zwei Arten von Fehlern bei Umfragen unterschieden werden: Stichprobenfehler, die die Repräsentativität mindern, und Messfehler, die während der Datenerhebung selbst die Datengüte mindern (Taddicken, 2008). Bei den Stichprobenfehlern kann weiterhin zwischen Coverage-, Sampling- und Nonresponse-Einschränkungen der Repräsentativität differenziert werden, während Messfehler durch den Befragten, den Interviewer, das Instrument und die Methode resultieren können (Kühle & van Ackeren, 2012). Bei reinen Webbefragungen sind Coverage-Stichprobenfehler zu befürchten, da möglicherweise nicht alle vorgesehenen Untersuchungseinheiten ausreichende technische Zugangsmöglichkeiten zur Webbefragung besitzen oder sie vorher nicht ausreichende technische Kenntnisse erworben haben. Diese Personen können dann nicht an der Befragung teilnehmen bzw. werden nicht durch das Befragungsinstrument erreicht und sind als Teil der Grundgesamtheit unterrepräsentiert (Undercoverage-Effekt) (Maier et al., 2012). Systematische Sampling-Fehler treten ggf. auf, wenn eine reine webbasierte Stichprobenziehung vorgenommen wird und die realisierte Stichprobe bzw. die Grundgesamtheit nicht ausreichend bekannt sind. Dies kann aber für die vorliegenden Studien ausgeschlossen werden. Nonresponse-Effekte treten als Unit-Nonresponse (vollständiges Ausbleiben von Antworten) und Item-Nonresponse auf. Letzterem kann bei Web-Befragungen unter Umständen dadurch begegnet werden, dass innerhalb der

Webbefragung ein Antwortzwang integriert wird, der die nächste Fragebogenseite erst nach vollständiger Beantwortung aller vorherigen Fragen öffnet. Dadurch steigt aber die Gefahr, dass die Befragung gänzlich abgebrochen wird (Block, Klein, van Ackeren & Kühn, 2011).

Durch die flächendeckende Ausstattung von Schulen in Nordrhein-Westfalen mit internetfähigen Computern und die technikaffine Zielgruppe[98] der Studien, sollten allein durch eine Webbefragung bedingte Undercoverage-Effekte ausgeschlossen werden können. Durch die Kombination der Befragungsmethoden sind aber Präferenzen ähnlich eines Coverageeffekts zu erwarten. Jüngere Lehrkräfte sind durchschnittlich mit Webbefragungen besser vertraut als ältere Lehrkräfte und werden daher überproportional die webbasierte Fragebogenversion ausfüllen, während sich dies bei älteren Lehrkräften umgekehrt verhalten sollte. Um gleichzeitig Overcoverage-Effekte auszuschließen, war auch die Teilnahme an der Webbefragung nur über eine Fragebogen-Identifikationsnummer möglich. Besonders die Verweigerung, an Webbefragungen teilzunehmen (Unit-Nonresponse), hätte sich hingegen als problematisch darstellen können, da mit der Internet-Affinität möglicherweise auch die Gefahren dieses Mediums bewusster sind. Studien zeigen außerdem, dass die Rücklaufquote für Webbefragungen häufig niedriger ist als bei postalischen Befragungen (van Ackeren, Block, Klein & Kühn, 2012). Die webbasierten Pilotierungsstudien stützten diese Annahme. Die Teilnahmequote lag jeweils unter zwanzig Prozent, sodass in den Pilotierungsstudien eine sehr hohe Zahl an Uni-Nonresponse beobachtet werden konnte. Auf eine völlig webbasierte Befragung wurde daher verzichtet.

Unter den Messfehlern sind für die Web-Befragung vor allem Instrumenteneffekte und Methodeneffekte bedeutsam. Instrumenteneffekte sind bei Webbefragungen vor allem als durch die Struktur der Fragen und Frageblöcke bedingte Effekte wahrscheinlich (Kühle & van Ackeren, 2012). Auch die Länge des Fragebogens kann die Ergebnisse insbesondere bei denjenigen Personen verzerren, die es nicht gewohnt sind, länger an Bildschirmen zu arbeiten (Taddicken, 2008). Um die Instrumenteneffekte möglichst gering zu halten, wurden der Fragebogen in der gedruckten Version und der Fragebogen in der webbasierten Version möglichst identisch gestaltet. Frageblöcke der gedruckten Version wurden nach Möglichkeit auch stets als ein Fragebogenblock in der Webbefragung präsentiert[99] und auf Pflichtfragen wurde ebenfalls verzichtet. Lediglich ein Filter der gedruckten Fragebogenversion bei Frage 8 wurde berücksichtigt und in der Onlineversion als automatischer Filter integriert. Auch die Darstellung wurde genauso schlicht gehalten wie dir gedruckte Version. Auf die Möglichkeit, den Rücklauf von Webbefragungen durch ein ansprechendes Design zu erhöhen, wurde verzichtet.

Bei den Methodeneffekten handelt es sich um Verzerrungen, die durch das Gefühl der Befragten entstehen, zu bestimmten Antworten gedrängt zu werden. Typischerweise liegt

[98] Die Zielgruppe sind Mathematiklehrkräfte. Die Nutzung von Computern ist verbindlicher Teil des Kernlehrplans für Mathematik in Nordrhein-Westfalen und der Erwerb der notwendigen Medienkompetenz dadurch für alle Personen der Zielgruppe obligatorisch.

[99] Beim Frageblock in Studie A zu den Unterrichtszielen war dies allerdings nicht möglich, da gleichzeitig kein Scrollen nötig sein sollte.

dieser Effekt vor, wenn Effekte der sozialen Erwünschtheit auftreten, welches besonders bei unangenehmen Fragen zu erwarten ist (Taddicken, 2008). Im Konzept der sozialen Erwünschtheit kann nach Musch, Brockhaus und Bröder (2002) vertiefend zwischen der Selbsttäuschung und der Fremdtäuschung unterschieden werden. Die Fremdtäuschung entspricht dabei der klassischen Vorstellung der sozialen Erwünschtheit, dass Befragte in der Weise antworten, mit ihrer Antwort *vor anderen* in einem besseren Licht dazustehen. Unter Selbsttäuschung wird eine Tendenz verstanden, sein Selbstbild *vor sich selbst* positiver erscheinen zu lassen (Musch, Brockhaus & Bröder, 2002). Taddicken (2008) vermutete mit Bezug zu Studien, die eine höhere private Selbstaufmerksamkeit bei Webbefragungen gegenüber postalischen Befragungen nachwiesen, dass die höhere private Selbstaufmerksamkeit zu geringeren Fremd- und Selbsttäuschungstendenzen bei Webbefragungen führt. Sie konnte in einer experimentalen Studie allerdings keine signifikanten Unterschiede zwischen webbasierten und postalischen Befragungen in der Fremdtäuschung belegen. Unterschiede zeigten sich hingegen derart, dass die Tendenz zur Selbsttäuschung in den Webbefragungen ihres Experiments höher war (Taddicken, 2008).

Die parallele Darbietung der Fragebögen der zugrunde liegenden Studien A und B auf postalischem und webbasiertem Weg kann zu Methodeneffekten führen, sofern die Fragebögen Items enthalten, die für eine Selbsttäuschungstendenz anfällig sind. Da nicht geklärt ist, ob evtl. Täuschungsabsichten durch Testcoaching einer Fremd- oder Selbsttäuschungsabsicht entspringen, können entsprechende Verzerrungen im Vorfeld nicht ausgeschlossen werden, sondern mussten nachträglich überprüft werden. Dazu wurden die für die Vorbereitung aufgewendeten Unterrichtsstunden und die fachdidaktische Selbstwirksamkeitsskala als Vergleichsmaßstab verwendet. In beiden Studien zeigte sich für drei der vier Altersgruppen kein signifikanter Unterschied im Vorbereitungsumfang zwischen Befragungsteilnehmenden, die postalisch teilnahmen, und denjenigen, die die webbasierte Version des Fragebogens beantworteten. In Studie A gab es auf Niveau p=0.01 für die Altersgruppe der „36-45 Jahre“ und in Studie B für die Altersgruppe der „56-65 Jahre“ einen signifikanten Unterschied mit n=81, t(17.92)=2.80, r=34 bzw. n=84, t(27.03)=2.91, r=14, wobei nur der erste Unterschied als mittlerer Effekt auffällt, der zweite hingegen nicht einmal ein schwacher Effekt ist. Beim Vergleich der Mittelwerte für die Skala „fachdidaktische Selbstwirksamkeitserwartung“ zeigten sich überhaupt keine signifikanten Unterschiede zwischen beiden Zugängen für die vier Gruppen. Beide Zugänge können folglich für diese beiden Studien als gleichwertig angesehen werden.

6.1.4 Design des Fragebogenpakets

Entsprechend den gesetzlichen Regelungen zur Befragung von Lehrkräften in Nordrhein-Westfalen müssen die Schulleitungen der Schulen um Erlaubnis gebeten werden, wenn man Lehrkräfte der jeweiligen Schule befragen möchte. Dies geschah durch zwei einseitige

Anschreiben. Den Anschreiben an die Schulleitungen wurden dabei gleich die Fragebogenpakete für die Lehrkräfte mit der Bitte beigelegt, diese an diejenigen Mathematiklehrkräfte ihrer Schule weiterzureichen, die im Sommerhalbjahr 2010 eine achte Klasse in Mathematik unterrichteten.

Die Fragebögen waren jeweils zehn Seiten lang. Auf den ersten fünf Seiten wurde die Vorbereitung auf die Vergleichsarbeiten im Schuljahr 2009/10 in beiden Studien identisch erfragt. Der zweite Teil im Fragebogen der Studie A umfasste Beurteilungsaufgaben zu den drei Überzeugungsbereichen des Lehrer-Handlungskompetenz-Modells nach Baumert und Kunter ohne den Bereich „Wissen und Können“. Außerdem decken die Beurteilungsaufgaben allgemeine Persönlichkeitsvariablen aus der Feedback-Interventions-Theorie nach Kluger und DeNisi ab. In Studie B folgten im zweiten Teil des Fragebogens Items mit direktem Bezug zu VERA 8 in Anlehnung an Studien der Rezeptionsforschung und waren ebenfalls geeignet, die Persönlichkeitsvariablen aus der Feedback-Interventions-Theorie abzubilden. Dabei wurde aber ein stärkerer Gegenstandsbezug zu VERA8 hergestellt. Neben dem Fragebogen sollte jede Lehrkraft zwei Anschreiben (eines der wissenschaftlichen Projektleitung, eines zur Betonung der Studien als Grundlage für das Promotionsvorhaben), einen Antwortumschlag sowie ein A6-Blatt für die Teilnahme an einem Gewinnspiel erhalten. Mittels des A6-Blattes wurden Kontaktdaten erfragt, die dadurch unabhängig von den Fragebogendaten erfasst werden konnten. Verlost wurden zwanzig Bücher-Gutscheine und zwanzig Bücher mit Unterrichtsanregungen. Allgemein wird angenommen, durch solche so genannten Incentives die Teilnahmebereitschaft zu erhöhen (Diekmann & Jann, 2000). Die Lehrkräfte wurden im Anschreiben gebeten, den Fragebogen innerhalb der nächsten 14 Tage zu beantworten. 28 Tage nach Versanddatum wurde an alle Schulen ein Brief verschickt, in dem den bisherigen Befragungsteilnehmern für ihre Beteiligung gedankt wurde und alle anderen erneut gebeten wurden, den Fragebogen noch zu beantworten.

6.1.5 Angestrebte Stichproben

Zu Beginn des Schuljahres 2009/2010 gab es in NRW laut amtlichem Adressverzeichnis 629 Gymnasien. Davon wurden dreizehn im Vorhinein ausgeschlossen. Die Gründe dafür waren die Teilnahme an einer der Pilotierungen, keine oder nur eine Klasse in der achten Jahrgangstufe oder eine Schulstruktur innerhalb der Schule, deren Ablauf nicht mit den anderen Gymnasien vergleichbar ist und eher Gesamtschulen ähnelt. Eine Schule wurde nachträglich ausgeschlossen, weil die Schule selbst angab, keine achte Jahrgangsstufe zu besitzen. Fünf Schulen wurden für die Studie B ausgewählt, um an ihnen mittels „Fragebogenpaten“ Vollerhebungen durchführen zu können.[100] Dazu wurden Lehrkräfte und

[100] Kunter und Klusmann vermuten, dass bei Kompetenzmessungen unter Lehrkräften vorwiegend Lehrkräfte teilnehmen könnten, die bzgl. der Kompetenz ein positives Selbstkonzept besitzen. Sie schlagen daher für die Repräsentativitätsprüfung der realisierten Stichprobe den Vergleich mit anderen repräsentativen Stichproben

Referendare gewonnen, die ihre Kollegen an die Beteiligung an der Studie erinnern sollten. Die verbleibenden 610 Gymnasien wurden erst nach Postleitzahl und anschließend alphabethisch sortiert und abwechselnd der ersten oder zweiten Studie zugeordnet und angeschrieben. Durch dieses Vorgehen entstanden theoretisch für beide Studien Klumpenstichproben, wobei die Klumpen zufällig gezogen wurden.

Die tatsächliche Zahl der Lehrkräfte, die an den 615 Gymnasien in der achten Jahrgangsstufe Mathematik unterrichtet, wurde auf Grundlage statistischer Kennzahlen geschätzt: Der Statistische Bericht für allgemeinbildende Schulen der Schuljahre 2007/08, 2008/09 und 2009/10 (Landesamt für Datenverarbeitung und Statistik, 2008, 2009, 2010) weist Schülerzahlen aus, nach denen durchschnittlich 112 Schüler und Schülerinnen in der achten Jahrgangsstufe im Schuljahr 2009/10 unterrichtet wurden. Dies entspricht vierzügigen Jahrgangsstufen und folglich vier Lehrkräften pro Gymnasium[101]. Für die meisten Schulen lag jeweils nur die Schülergesamtzahl aller Jahrgangsstufen vor, nicht aber jahrgangsstufenweise. Die Jahrgangsstufenzügigkeit lag nur bei ca. zehn Prozent der Schulen vor. Unter Zurhilfenahme der Werte dieser 60 Gymnasien und den jeweiligen Schülerzahlen wurde für die verbleibenden 555 Gymnasien die Zahl der Mathematiklehrkräfte in der jeweiligen achten Jahrgangsstufe geschätzt. Zum Einsatz kam dabei für Schulen mit bis zu 1400 Schülern und Schülerinnen ein Maximum-Likelihood-Verfahren, bei dem die für eine Gesamtschülerzahl höchste bekannte Jahrgangstufenzahl angenommen wurde. Dieses Verfahren senkt die Wahrscheinlichkeit, zu wenig Lehrkräfte für eine Schule anzunehmen, erhöht aber zusammen mit der Möglichkeit, dass eine Lehrkraft mehrere achte Klassen unterrichtet, die Wahrscheinlichkeit, die tatsächliche Anzahl an Lehrkräften zu überschätzen. Für größere Schulen (mit mehr als 1400 Schülerinnen und Schülern) wurde die Zügigkeit über eine durchschnittliche Klassengröße von 30 Schülerinnen und Schülern geschätzt. Eine theoretisch mögliche Unterschätzung der Jahrgangszügigkeit scheint gerechtfertigt, da hier vermehrt Lehrkräfte mehrere Parallelklassen unterrichten. Aufgrund dieser Schätzverfahren wurden in Studie A 1283 und in Studie B 1300 Fragebögen verschickt.

Als Kontrolle für diese beiden Verfahren wurden alle teilnehmenden Lehrkräfte gebeten, die Anzahl der in der achten Jahrgangsstufe Mathematik unterrichtenden Lehrkräfte anzugeben. Mittels dieser Zahl ist es außerdem möglich, in Kombination mit den fortlaufenden Fragebogennummern eine relativ zuverlässige Zuordnung einzelner Fragebögen zur selben Schule vorzunehmen, ohne dass dabei die Anonymität der Schulen aufgehoben wird.

Nach Garner (2005) lässt sich der Rücklauf bei postalischen Befragungen steigern, wenn auf den Fragebögen ein Haftnotizzettel mit handschriftlicher Ausfüllbitte aufgeklebt ist. Das Verfahren wurde für beide Studien übernommen. Da bisher aber nur

---

vor. Durch den Vergleich soll analysiert werden, ob die realisierte Stichprobe bzgl. für die Fragestellung relevanter Merkmale abweicht (Kunter & Klusmann, 2010). Durch Fragebogenpaten soll diese Vergleichsstichprobe gewonnen werden. Ursprünglich sollten es mindestens zehn Schulen sein. Es haben sich aber nur fünf Referendare bzw. Lehrkräfte als Fragebogenpaten engagiert.

[101] Aus sprachlicher Verknappung wird bei Schulen mit n Klassen in der achten Jahrgangsstufe, die von y Lehrkräften in Mathematik unterrichtet werden von y-zügigen Schulen gesprochen.

Untersuchungsergebnisse für Hochschulangehörige vorlagen (Garner, 2005; Kreutz, Hahn & van Ackeren, 2011), wurde nur ca. die Hälfte der Fragebögen mit einem entsprechenden Haftnotizzettel versehen (in Studie A 621 und in Studie B 576 der Fragebögen). Die handschriftliche Notiz lautete „Wären Sie bereit, dies auszufüllen? Vielen Dank! ☺“. Mit einer solchen Notiz versehene Fragebögen werden signifikant häufiger ausgefüllt (Garner, 2005). Dabei steigt die Bereitschaft, an einer schriftlichen Befragung teilzunehmen, ohne dass sich die zusätzlich Teilnehmenden wesentlich bzgl. allgemeiner Persönlichkeitsmerkmale unterscheiden (Kreutz et al., 2011). Dies muss aber nicht gleichermaßen für das offensichtliche Thema der Befragung, z.B. in diesem Fall das Vorbereitungsverhalten gelten. Für den deskriptiven Teil (Forschungsfrage 1) wird daher zwischen Fragebögen mit und ohne Haftnotizzettel unterschieden. Aus Studie A werden dadurch Studie A1 (mit Haftnotiz) und Studie A2 (ohne Haftnotiz) sowie aus Studie B Studie B1 und Studie B2.

Aufgrund der Weitergabe der Fragebogenmaterialien durch die Schulleiter an die Lehrkräfte kann es neben dem im Vorhinein vorgenommenen Ausschluss zum weiteren Ausschluss ganzer Schulen durch die Schulleitungen kommen. Neben Teilnahmeverweigerung durch einzelne Lehrkräfte entstand dadurch an zwei Stellen eine Selbstselektion (Nonresponse-Fehler und Undercoverage-Effekt). Insgesamt entstand somit eine Ad-hoc-Stichprobe (Weick, 1976). Da weder die Rücklaufquote auf Schulebene[102] noch auf Lehrerebene[103] eindeutig vorliegen und beide in die vorliegende Gesamtquote eingehen, können nur Mindestangaben zur Beteiligungsquote auf Schul- und Lehrerebene gemacht werden.

[102] Der zu erwartende Rücklauf kann für jede einzelne Schule theoretisch mittels einer Binominalverteilung geschätzt werden. Voraussetzung hierfür ist allerdings die Kenntnis, mit welcher Wahrscheinlichkeit eine einzelne Lehrkraft an der Studie bereit ist teilzunehmen. Diese Wahrscheinlichkeit müsste für jede Einzelschule vorliegen, aber schon aggregierte Werte schwanken beträchtlich. Kühle berichtet beispielsweise von Rücklaufquoten zwischen 34,4% für Vorsitzende der Fachkonferenz Mathematik und 39% der Mathematiklehrkräfte insgesamt (Kühle, 2010), Maier von Beteiligungsquoten von 18,1% und 42,6% bei Befragungen an Gymnasien. Alternativ kann die Wahrscheinlichkeit der Beteiligung einer Lehrkraft auf Lehrerebene auch auf Basis der maximal möglichen Untersuchungsteilnehmenden und der tatsächlichen Teilnehmenden geschätzt werden. Ignoriert wird in dem Fall, dass Schulleitungen die Fragebögen eventuell nicht weiterleiteten, Lehrkräfte dieser Schulen sich aber beteiligt hätten. Beide Verfahren sind daher nicht zuverlässig zu realisieren.

[103] Auf Lehrerebene kann die Beteiligungsquote aus der Kombination von ausgefüllten Fragebögen, maximal möglichen Untersuchungsteilnehmenden und der angenommenen Wahrscheinlichkeit für die Weitergabe der Fragebögen durch die Schulleitung geschätzt werden. Verlässliche Daten dazu liegen aber nicht vor. So schwankte die Schulbeteiligung in zwei vergleichbaren Studien von Kühle zwischen 79% und 65% (Kühle, 2010).

6.1.6 Realisierte Stichproben

Insgesamt liegen in Studie A n=380 Fragebögen von 209 verschiedenen Gymnasien vor. In Studie B sind es n=317 von 180 Gymnasien. Für die Rücklaufquote auf Lehrerebene[104] ergeben sich 34,7% (Studie A)[105] und 29,1% (Studie B), auf Schulebene[106] 68,5% und 58,9%. Vollständig haben in Studie A 31 Gymnasien und in Studie B 32 Gymnasien teilgenommen. Die Rücklaufquote liegt für beide Studien damit im selben Rahmen wie die Rücklaufquote vergleichbarer Studien. Verallgemeinernde Schlüsse auf die Population der entsprechenden Mathematiklehrkräfte insgesamt sind dadurch kritisch zu hinterfragen. Es ist aber durchaus üblich, Ergebnisse bei dieser Rücklaufquote als repräsentativ zu betrachten (z.B. Kühle 2010, Maier, 2009). Die Rücklaufquote der Schulen mit Fragebogenpaten beträgt 52,3%. Das hier angestrebte Ziel eines vollständigen Rücklaufs wurde folglich verfehlt, sodass die Ergebnisse nicht als Referenzabgleich taugen.

Für das Schuljahr 2009/10 liegen die Alters- und Geschlechterverteilung der Mathematiklehrkräfte an Gymnasien vor. Wie Tab. 6.1 zu entnehmen ist, wiesen die realisierten Stichproben große Verzerrungen bzgl. der Alters- und Geschlechterstruktur im Vergleich zur Gesamtverteilung der Mathematiklehrkräfte an Gymnasien in NRW auf. Frauen waren in beiden Stichproben deutlich unterrepräsentiert. Die Geschlechterverteilung muss allerdings nicht für einzelne Jahrgangsstufen gelten. Frauen nutzen häufiger die Möglichkeit in Teilzeit zu arbeiten. Dadurch sinkt die Wahrscheinlichkeit, dass eine weibliche Lehrkraft auch tatsächlich in der achten Jahrgangsstufe eingesetzt wird. Auch die Altersverteilung wich für beide Studien von der Altersverteilung insgesamt ab. Dies lässt sich durch die Leitungsstrategie erklären, in der Mittelstufe vermehrt junge, „unverbrauchte“ Lehrkräfte (häufig auch Referendarinnen und Referendare) einzusetzen, die den höheren Belastungen durch die Schülerinnen und Schüler mehr gewachsen scheinen und über die geringere Altersdifferenz zu den jenen eine größere Vorbildfunktion besitzen. Die Abweichungen zwischen den Stichproben und den statistischen Daten der Gesamtpopulation sind folglich

[104] Nimmt man an, dass Ausfälle beim Rücklauf vorwiegend durch nicht weitergeleitete Fragebögen entstanden sind, erhöhen sich die Werte auf 50,7% (Studie A) und 49,3% (Studie B). Diese Werte können nicht verifiziert werden, zeigen aber den möglichen Einfluss der Schulleitung auf die Rücklaufquote. Bedenkt man, dass im Fragebogen der Studie B mehrfach nach Verhalten der Schulleitung gefragt wird, scheint die Schulleitung als Fehlerquelle nachvollziehbar.

[105] Die angegebenen Rücklaufquoten berücksichtigen, dass mehr Fragebögen versendet wurden als Lehrkräfte in der Zielgruppe waren. Die Grundgesamtheit der Lehrkräfte wurde dabei unter Kenntnis der Angaben aus den Fragebögen geschätzt, sodass eine genauere Schätzung der Grundgesamtheit vorgenommen werden konnte als dies vor der Erhebung möglich war. Diese Schätzung kann nicht genutzt werden, um die Rücklaufquote für Fragebögen mit und ohne Haftnotizzettel anzugeben. Bei einer konservativen Schätzung der Grundgesamtheit betragen die Rücklaufquoten für die Studie A1 33,5%, für die Studie A2 26,4%, für die Studie B1 28,5% und für die Studie B2 18,9%.

[106] Die geschätzte Teilnehmerquote beträgt auf Schuleben für Studie A 87,4% (angenommene Lehrerrücklaufquote von 34,7%) bzw. 84,9% (angenommene Lehrerrücklaufquote von 29,1%) bei Studie B, wenn man Schulen ergänzt, an denen die Fragebögen weitergeleitet wurden, sich aber keine Lehrkraft beteiligen wollten. Sowohl die korrigierten Schätzung für die Beteiligungsquote auf Lehrer- als auch auf Schulebene nehmen die stochastische Unabhängigkeit der Teilnahmeverweigerung auf beiden Ebenen an.

nachvollziehbar. Trotzdem wird durch diese plausible Erklärung nicht ausgeschlossen, dass sich in den Abweichungen zwischen realisierter Stichprobe und Gesamtpopulation der Mathematiklehrkräfte systematische Verzerrungen verbergen.

Tabelle 6.1

Struktur der realisierten Stichproben und Gesamtverteilung von Mathematiklehrkräften an Gymnasien in NRW

	NRW		Studie A		Studie B	
	insgesamt	Frauen	insgesamt	Frauen	Insgesamt	Frauen
Altersstufen	(%)	(%)	(%)	(%)	(%)	(%)
insgesamt		60.4%		46.5%		39.3%
<36	17.2%	13.3%	28.8%	37.5%	28.1%	42.3%
36-45	21.4%	20.9%	21.9%	17.9%	20.4%	17.1%
46-55	21.8%	17.8%	20.8%	24.4%	23.0%	25.2%
>55	39.5%	47.9%	28.5%	20.0%	28.4%	15.4%

Über die Frage „*Wie viele Lehrkräfte unterrichten in diesem Halbjahr in der achten Jahrgangsstufe Ihrer Schule Mathematik?*“ lässt sich überprüfen, ob aufgrund der Verfahren zur Schätzung der Jahrgangszügigkeit ausreichend viele Fragebögen an die Schulen versendet wurden.[107] Rechnet man die Angaben der Lehrkräfte auf alle Gymnasien hoch, hätten in der Studie A 1095 Fragebögen verschickt werden müssen, in Studie B wären 1091 Fragebögen ausreichend gewesen. Eine Analyse für die einzelnen Schulgrößen ergibt, dass wesentlich mehr achte Jahrgangsstufen von nur zwei oder drei Lehrkräften unterrichtet werden und zu häufig vier und fünf Lehrkräfte in dieser Jahrgangsstufe vermutet wurden. Nur an einem Gymnasium sind erkennbar zu wenige Fragebögen angekommen, nach

[107] Maier folgert aus Differenzen zwischen den von ihm aus der amtlichen Statistik berechneten durchschnittlichen Klassengrößen und den Angaben der Lehrkräfte in seiner Stichprobe eine Stichprobenverzerrung. Dies wäre auch hier eine mögliche Interpretation. Es scheint aber weder einen sinnvollen Grund dafür zu geben, dass eher Lehrkräfte teilnehmen, die mit wenigen Kolleginnen und Kollegen in der Jahrgangsstufe unterrichten, noch für den Befund in Maiers Untersuchung, in denen Hauptschul- und Gymnasiallehrkräfte in kleineren Klassen weniger belastet sind als in großen Klassen und daher häufiger antworten, dies aber auf Realschullehrkräfte nicht zutreffen soll (Maier, 2009). Gegen Maiers Interpretation spricht, dass er die Größen der für seine Studie relevanten sechsten Jahrgangsstufen für Hauptschulen und noch mehr für Gymnasien deutlich überschätzt hat. Gegen eine Stichprobenverzerrung dieser Studien spricht, dass mehr Schulen mit „zweizügigen Jahrgangsstufen“ sich beteiligten als überhaupt nach der Schätzung existieren sollten.

Meldung durch den Mittelstufenkoordinator konnte dieses Problem aber noch behoben werden.

6.2 Fragebogen-Items[108]

Im folgenden Abschnitt werden der Fragebogen aus Studie A und der Fragebogen aus Studie B vorgestellt. Da die erste Hälfte der beiden Fragebögen identisch war, wird erst später eine Differenzierung nach Studien vorgenommen. Die erste Fragebogenhälfte wird vollständig dargestellt. Die zweiten Hälften der Fragebögen werden soweit beschrieben, wie sie Aufnahme in die Analysen und Modelle (vgl. 6.3) dieser Arbeit gefunden haben. Die Erläuterung an dieser Stelle umfasst daher nicht diejenigen Items, die keine Aufnahme in die Auswertung fanden und deren Datenergebnis für spätere Auswertungen vorgesehen ist. Die Antwortalternativen einiger Items sind allerdings erst im nächsten Kapitel innerhalb der Ergebnisdarstellung aufgeführt. Bei etablierten Skalen wird lediglich eine Begründung für die Auswahl gegeben, sodass für die genauere Beschreibung auf die ursprünglichen Quellen verwiesen werden muss.

6.2.1 Items zum Bereich Testcoaching

Die Items zum Bereich Testcoaching entstanden aus den Leitfragen der vorab durchgeführten qualitativen Interviewstudie und den von den Lehrkräften darin geäußerten Antworten. Die Antworten der Lehrkräfte in der Interviewstudie dienen als Referenzrahmen.

a) Der zeitlicher Vorbereitungsumfang

Der zeitliche Vorbereitungsumfang wurde als offenes Feld in ungefährer Unterrichtsstundenzahl erfragt (Frage 1). Von einer Auswahlfrage mit Einteilung in Schulwochen wurde abgesehen, da nicht eindeutig ist, wie viele Stunden Mathematik tatsächlich in den jeweiligen Klassen pro Woche unterrichtet werden. Vorgesehen sind drei bis vier Wochenstunden. Möglich wäre auch eine Auswahlfrage mit Einteilung nach den Wirkungsstufen gewesen, die Bunting und Mooney (2001), Messick und Jungeblut (1981) sowie Flippo, Becker und Wark (2000) in ihren (Meta-)Analysen zur Wirkung von Testcoaching berichten. Dadurch wären aber größere Vorbereitungsumfänge nicht sinnvoll

[108] Die verwendeten Instrumente aus Eigenentwicklung sind im Anhang zu finden.

abbildbar gewesen. Aufgrund der Interviews waren hier Werte zwischen null und zwanzig zu erwarten, Wirkungsunterschiede wurden bisher aber nur bis maximal zwölf Stunden untersucht.

b) Die Qualität der unmittelbaren Vorbereitung

Die Qualität der unmittelbaren Vorbereitung wurde durch drei Fragen erhoben. Frage 2 beinhaltete Items zum Einsatz von (VERA-ähnlichen) Aufgaben in Unterricht und vorherigen Klassenarbeiten sowie zur Nutzung der Informationsquellen zu den Vergleichsarbeiten im Internet o.ä. Maßnahmen, die als Familiarity Approach klassifiziert werden können. Es wurde erwartet, dass Lehrkräfte alte Testaufgaben sehr häufig zu Übungszwecken eingesetzt haben, alle anderen Optionen hingegen eher selten angegeben werden. Außerdem wurde in Frage 2 abgefragt, ob bestimmte Inhaltsbereiche oder Prozesskompetenzen besonders geübt oder den Schülern zu üben nahe gelegt wurden. Dies kann als Content Approach eingestuft werden. In den Interviews äußerten fast alle Lehrkräfte, bestimmte Prozess- oder Inhaltskompetenzen aus den Kernlehrplänen NRW gezielt geübt zu haben. Da es sich damals aber auch um Deutsch- und Englischlehrkräfte handelte und seit 2009 in Mathematik keine Schwerpunkte mehr bei den Prozess-Kompetenzen gelegt werden, sondern die Bildungsstandards vollständig als Referenz dienen, sollte ein anderes Vorbereitungsverhalten erfolgen. Schwerpunktsetzungen sind dann nicht mehr durch das Ziel zu erklären, Testergebnisse zu beschönigen.

Die konkrete Fragestellung lautete: „Welche der folgenden Optionen haben Sie in dieser Klasse als unmittelbare Vorbereitung auf die LSE genutzt? (Mehrfachnennungen möglich) - Ich habe...“. Die entsprechenden Optionen (vgl. nächstes Kapitel) sollten angekreuzt werden. Die zweiwertige Antwortskala lässt es außerdem zu, einen Summenindex der durchgeführten Maßnahmen zu bilden. Dadurch wurde mit Frage 2 auch eine Form des Vorbereitungs*umfangs* erhoben. Zusätzlich gab es anschließend ein Feld für freie Antworten. Durch die vorgegebene Liste konnten dem Untersuchungsgegenstand entsprechend zielführende Antworten erreicht werden. Gleichzeitig sank aber die Wahrscheinlichkeit für die Angabe von zusätzlich frei formulierten Optionen.

Ebenfalls zu Familiarity Approach gehörten die in Frage 3 zum Vertrautmachen mit den Aufgabenformaten erfragten Unterrichtsinhalte. Hier lautete die Frage: „Welche der folgenden Themen haben Sie im Unterricht angesprochen? - Ich habe angesprochen...“. Abgefragt wurden die Themen „das richtige Verstehen der Aufgabenstellung“, „die Entnahme der wichtigen Informationen aus der Aufgabenstellung“ und „wie man im Hinblick auf die Antwortformate richtig antwortet“. Als Antwortalternativen standen „mehrfach“, „einmal“ und „gar nicht“ zur Auswahl. Für die absolute Häufigkeit der Antwortverteilungen konnte aus den Interviews keine Prognose abgeleitet werden. Aufgrund der höheren Komplexität sollte aber das Thema „Entnahme der wichtigen Informationen aus der Aufgabenstellung“ seltener angesprochen worden sein (vgl. nächstes Kapitel).

Frage 6 umfasst eine Liste mit neun Test-Wiseness-Strategien. Die Strategien wurden zum Teil der Übersicht von Flippo, Becker und Wark (2000) entnommen und stammten zum Teil aus Interviewaussagen der Teilnehmenden aus der Vorstudie. Interviewte Lehrkräfte gaben 2008 allerdings an, auf solche Strategien nur sehr begrenzt Wert gelegt zu haben. Die Lehrkräfte sollten angeben, ob sie ihren Schülerinnen und Schülern diese Strategie mit auf den Weg gegeben haben bzw. ob sie ihnen zu einer Strategie geraten haben. Zusätzlich gab es die Möglichkeit selbst Strategien zu nennen, die nicht Teil der Liste waren.

c) Der Einsatz von Vorbereitungsheften und Lernhilfen

Von verschiedenen Schulbuchverlagen gibt es speziell für die Vorbereitung auf VERA8 (und VERA3) konzipierte Vorbereitungshefte, die jeweils eine Kompaktzusammenfassung der mathematischen Teilkompetenzen sowie Beispielaufgaben beinhalten. Diese richten sich zum Teil direkt an die Schüler und Schülerinnen und sind für zusätzliche Übungsphasen außerhalb der Schule nutzbar, sie können aber auch zur Vorbereitung im Unterricht eingesetzt werden. An Schulen wurden von den verschiedenen Anbietern dieser Hefte Gratisexemplare verschickt, sodass offensichtlich ein flächendeckender Einsatz angestrebt wurde. Ähnlich verhält es sich mit als „Lernhilfe“ bezeichneten Heften, in denen die erwarteten Kompetenzen zusammengefasst sind. Diese sind prinzipiell nicht auf ein Testergebnis ausgelegt, sondern dienen als persönliche Standortbestimmung. Den Interviews folgend war ein Einsatz der Hefte (vorwiegend Vorbereitungshefte) in mehr als fünfzig Prozent der Klassen zu erwarten. Gefragt wurde, ob solche Hefte (getrennt nach Vorbereitungsheften und „Lernhilfen“) *in (fast) allen Stunden*, *in einigen Stunden* oder *gar nicht* eingesetzt wurden. Weiter wurde gefragt, ob mit diesen Vorbereitungsheften alle Inhaltskompetenzen und ob mit ihnen alle oder einige Prozesskompetenzen wiederholt werden sollten. Zusätzlich sollten weitere Lehrkräfte die Anschaffung zumindest einzelnen Schülern nahe gelegt haben. Auch dies wurde in Frage 4 und Frage 5 erhoben.

d) Die Veränderung der Vorbereitung

Die Frage 8 diente zur Erfassung, ob die Vorbereitung sich im Laufe der Jahre bei einzelnen Lehrkräften geändert hat und welche Gründe es dafür ggf. gab. Die Antwortmöglichkeiten differenzierten zwischen einer *intensiveren* Vorbereitung, *gleich bleibender Intensität*, *geringerer Intensität* im Vergleich zum ersten Mal einer Vorbereitung auf VERA8/Lernstand8/Lernstand9 und einer *erstmaligen Konfrontation* mit zentralen Vergleichsarbeiten. In den Interviews hatten sich die ersten drei möglichen Tendenzen relativ gleichverteilt gezeigt. Lehrkräfte, die erstmalig mit zentralen Vergleichsarbeiten konfrontiert waren, gab es dort nicht.

Frage 9 [109] richtete sich ausschließlich an diejenigen Lehrkräfte, die bereits vorher Erfahrungen mit VERA8 gesammelt hatten und ihrer Ansicht nach dieses Mal intensiver oder weniger intensiv vorbereiteten als beim ersten Mal. Diese Lehrkräfte wurden gebeten anzugeben, warum sie die Intensität der Vorbereitung veränderten. Zu Auswahl standen jeweils drei vorformulierte Antwortmöglichkeiten. Bei höherer Intensität waren dies: Heute kenne ich mich mit den LSE besser aus./In den vergangenen Jahren hat der Druck für mich zugenommen, bei den LSE gut abzuschneiden./In den vergangenen Jahren hat der Druck für Schüler zugenommen, bei den LSE gut abzuschneiden. bzw. bei niedrigerer Intensität: Heute kenne ich mich mit den LSE besser aus./In den vergangenen Jahren hat der Druck für mich abgenommen, bei den LSE gut abzuschneiden./In den vergangenen Jahren hat der Druck für Schüler abgenommen, bei den LSE gut abzuschneiden. Außerdem konnten Gründe frei formuliert werden.

e) Die schuljahrumfassende Vorbereitung

Neben der unmittelbaren Vorbereitung, die eindeutig als Testcoaching identifiziert werden kann, könnte eine bestimmte Art der defizitminimierend-orientierten Unterrichtsgestaltung durchgeführt worden sein, die aber ebenfalls im Zusammenhang mit den Vergleichsarbeiten stehen kann. Dies betrifft insbesondere die Schwerpunktsetzung im achten Schuljahr insgesamt. Die Schwerpunktsetzung kann sowohl individuell auf eine Klasse abgestimmt als auch jahrgangsstufenumfassend erfolgt sein. In den Interviews gaben einige Lehrkräfte an, von Schuljahresbeginn an einzelne Teilkompetenzen besonders thematisiert und geübt zu haben, die sie als Testschwerpunkt der nächsten VERA erwarteten. In Frage 10 wurde erhoben, ob dieses Verhalten weiterhin auftritt und einzelne Teilkompetenzen besonders in den Vordergrund gestellt wurden oder hintenangestellt wurden. Dabei können dieses Mal rational nur vergangene Testergebnisse ausschlaggebend sein.

Gefragt wurde, ob die Lehrkraft für sich allein eine Prozesskompetenz intensiver oder weniger intensiv im aktuellen Schuljahr unterrichtete und ob es generell zu Kompetenzen analoge Absprachen in der Fachgruppe Mathematik gegeben hat. Zusätzlich sollten entsprechend Lehrkräfte angeben, ob dies aus ihrer Sicht mit Blick auf VERA8 geschehen sei.

f) Die außerunterrichtliche Vorbereitung

Möglich ist auch, dass eine Vorbereitung auf VERA8 außerhalb des Unterrichts durchgeführt wird. Dies kann im Rahmen von Übungsphasen sein, die von die Schülerinnen und Schüler individuell selbstständig organisiert sind, während eines Nachhilfeunterrichts oder auch von der Lehrkraft durch zusätzliche Übungsaufgaben unterstützt. Auch die Vorbereitungshefte und Lernhilfen können in diesem Zusammenhang zum Einsatz kommen. Vereinzelt äußerten

[109] Dieses Item findet in der Ergebnisdarstellung keine Berücksichtigung.

Lehrkräfte in den Interviews, ihren Schülerinnen und Schülern Übungsmaterialien für das selbstständige Üben zusammengestellt oder ihnen zur Nutzung von Nachhilfe geraten zu haben. Einigen Lehrkräften sollte es allerdings auch nicht bekannt sein, ob ihre Schüler und Schülerinnen sich zu Hause vorbereiteten.

In Frage 10 wurde daher weiterhin erhoben, ob Lehrkräfte einzelnen Schülern und Schülerinnen oder der ganzen Klasse empfohlen haben, sich speziell auf VERA8 vorzubereiten, ob Lehrkräfte einzelnen Schülerinnen und Schülern empfohlen haben, mit Blick auf VERA8 Nachhilfe zu nehmen, ob den Lehrkräften bekannt war, dass einige Schüler und Schülerinnen mit Blick auf VERA8 Nachhilfe genommen haben und ob dieser nach Einschätzung der Lehrkraft für VERA8 besonders übten.

g) Die Bedeutung der Vergleichsarbeiten für beteiligte Personen aus dem Schulumfeld

Die Lehrkräfte wurden nach einer Einschätzung gefragt, wie wichtig VERA8 für die Schüler und Schülerinnen (mehrheitlich), deren Eltern, für die Schulleitung und die anderen Deutsch-, Englisch- und Mathematiklehrkräfte der achten Jahrgangsstufe wohl gewesen sind. Ebenfalls sollten die Lehrkräfte auf der sechsstufigen Antwortskala der Frage 7 die persönliche Bedeutung von VERA8 angeben. Da Lehrkräfte sich möglicherweise nicht im Besitz der nötigen Information für eine solche Einschätzung sahen, gab es auch eine Antwortmöglichkeit „weiß nicht“. Die Interviews hatten gezeigt, dass dies häufiger der Fall sein kann. Grundsätzlich ließen die Interviews für Schülerinnen und Schüler, deren Eltern und die Schulleitung eine Einschätzung erwarten, die im Vergleich zu der persönlichen Bedeutung höher ausfallen sollte.

h) Die Funktionen von VERA8[110]

Während der Interviews zeigten sich unterschiedliche Vorstellungen über die Funktionen VERA8 bei den teilnehmenden Lehrkräften. Lehrkräfte schrieben zentralen Lernstandserhebungen andere Funktionen zu, als diese aus wissenschaftlicher Sicht zentralen Lernstandserhebungen zugeordnet werden. In Frage 11 wurden neun solcher „Funktionen“ vorgelegt, zu denen der Grad der Zustimmung oder Ablehnung auf einer sechsstufigen Skala angegeben werden sollte. Die neun Funktionen stehen erst einmal für sich, können aber zu Gruppen klassifiziert werden, wenn grundlegende Konzepte bei der Zuweisung unterstellt werden. Zusammen mit der VERA8 zugewiesenen Bedeutung ergibt sich somit eine Möglichkeit, die Wahrnehmung von VERA8 aus einer neuen Perspektive zu erheben. Fünf der neun Funktionen lassen sich klassifizieren, indem sie der Rechenschaftslegung oder der Qualitätsentwicklung zuordnen werden. Der Rechenschaftslegung wurden zugeordnet: „Arbeitszeugnis für Lehrkräfte“ und „Erstellen

[110] Dieses Item findet in der Ergebnisdarstellung keine Berücksichtigung und sind für eine spätere Veröffentlichung vorgesehen.

eines Schulrankings“. Die Perspektive der Qualitätsentwicklung drückt sich in folgenden Items aus: „Lehrkräften Hinweise für die weitere Unterrichtsplanung zu geben“, „…Lehrkräften Hinweise für die Klassenarbeitskonstruktion zu geben“ und „…den Schülern eine Rückmeldung über ihre mathematischen Kompetenzen zu geben“. Die Zuordnung/Struktur ist allerdings bisher erst aus Überlegungen hergeleitet und bedarf noch einer empirischen Überprüfung.

Aber auch die Frage nach der Leistungsattribution kann als Klassifikationsmerkmal angenommen werden. Hier können die Ergebnisse der Schülerinnen und Schüler bei VERA8 durch die Lehrkraft bei diesen verortet werden oder auf die Unterrichtsqualität zurückgeführt werden oder auf das Schulsystem als komplexes System. Die aus den Interviews gewonnen Funktionen lassen aber diesbzgl. keine ähnlich klare Struktur erkennen, wie man sie für das Begriffspaar „Qualitätsentwicklung/Rechenschaftslegung“ vornehmen kann. Dadurch ergeben sich vier nicht klassifizierte Funktionen, nämlich: „die Lehrer zu unterstützen, einfacher zu einer Zeugnisnote zu gelangen“, „eine Rückmeldung über unser Schulsystem zu geben“, „...für die Schüler eine Generalprobe für zentrale Prüfungen in der weiteren Schulzeit zu sein“ und „finanzielle und personelle Mittel an Schulen richtig zuzuweisen“.

i) Gründe für eine Vorbereitung und Gründe gegen eine Vorbereitung[111]

Im Theorieteil dieser Arbeit ist herausgearbeitet worden, dass eine Vorbereitung auf zentrale Vergleichsarbeiten je nach Qualität sinnvoll oder verzerrend sein kann. Auch die Überlegung, nicht vorzubereiten, kann sowohl aus dem Gedanken resultieren, der Anlage des Instruments dadurch am besten zu entsprechen, als auch der Überlegung zugrunde liegen, bewusst eine Verzerrung der Testwerte anzustreben. Diese Differenzierung kann theoretisch an der Vorbereitungsqualität gemessen werden, setzt aber bei den Lehrkräften das Bewusstsein darüber voraus, welche Art von Vorbereitung dem angestrebten Ziel am besten gerecht wird. In den Interviews haben sich Lehrkräfte mehrfach derart geäußert, dass dieses Wissen nicht vorausgesetzt werden kann. Daher wurden 12 Gründe für eine Vorbereitung auf VERA8 (Frage 12) und alternativ sechs Gründe gegen eine Vorbereitung (Frage 13) als Zustimmungs-Ablehnungs-Skala formuliert. Die Lehrkräfte sollten selbst entscheiden, ob sie vorbereitet haben und daher die zwölf Aussagen für eine Vorbereitung bewerten oder ob sie nicht vorbereitet haben und somit die anderen sechs Aussagen bewerten.

[111] Die beiden Itemblöcke wurde in den Fragebogen beider Versionen aufgenommen, die Ergebnisse aus den dadurch gewonnenen Daten werden aber schlussendlich nicht in den Analysen dieser Arbeit berücksichtigt, sondern ggf. an anderer Stelle veröffentlicht, da eine Motivsuche nicht im Fokus dieser Arbeit steht.

6.2.2 Skalen zum Lehrer-Handlungsmodell in Studie A

Die drei in Studie A abgebildeten Bereiche des Modells der Lehrer-Handlungskompetenz wurden konkret durch folgende drei Teildomänen abgebildet: (a) Transmitter-/Konstruktionsüberzeugungen [TÜ/KÜ], (b) personenbezogen Überzeugungen im Kontext des beruflichen Beanspruchungserlebens (UWES, BEL, OC, GW[112]) und (c) fachdidaktisches Selbstkonzept und fachdidaktische Selbstwirksamkeitserwartung im Mathematikunterricht (SK, SWE). Die Unterrichtsziele waren nicht integriert, obwohl sie Teil des Modells der Lehrer-Handlungskompetenz nach Baumert und Kunter sind. Die Gewichtung der Unterrichtsziele wird in dieser Arbeit erst in den Modellen M21 und M22 (s. u.) einbezogen.

Mit Ausnahme der hier gemessenen Selbstwirksamkeitserwartung und des fachdidaktischen Selbstkonzepts wurden jeweils (mehrfach) erprobte Skalen verwendet, die allerdings durch zusätzliche Items ergänzt wurden. Der Fragebogen in Studie A enthielt zusätzlich eine Skala zur Messung des Anstrengungsvermeidungsmotivs bei Lehrkräften, die von Wolfram Rollet entwickelt wurde. Die Items wurden in die Itemblöcke zu personenbezogen Überzeugungen im Kontext des beruflichen Beanspruchungserlebens eingeflochten.[113]

Alle Skalen wurden als sechsstufige Skalen mit Randlabels erfasst. Bei den Itemblöcken zur Leistungsattribution und den fachlichen Unterrichtszielen im Mathematikunterricht (siehe nächsten Abschnitt) wurden die Endpunkte mit „trifft gar nicht zu“ und „trifft voll zu“ betitelt, sodass es sich hier um unipolare Skalen handelt. Die Itemblöcke zu Transmitter-/Konstruktionsüberzeugungen, zum fachdidaktischen Selbstkonzept und der fachdidaktischen Selbstwirksamkeitserwartung im Mathematikunterricht, zu personenbezogen Überzeugungen im Kontext des beruflichen Beanspruchungserlebens und zur Gewissenhaftigkeit (GW - s. 6.3.3) wurden bipolar mit den Randlabels „lehne völlig ab“ und „stimme völlig zu“ eingesetzt.

a) Transmitter- und Konstruktionsüberzeugungen

Die Transmitter- und Konstruktionsüberzeugungen der Lehrkräfte wurden durch fünf bzw. sechs Items umfassende Skalen gemessen, die in MT21 für TEDS-M pilotiert wurden (Blömeke & Müller et al., 2008). Es handelt sich in ihrer Anlage um zwei unabhängige Skalen, die aber als theoretisches Konstrukt eindimensional gegenüber gestellt sind.

[112] Die Gewissenhaftigkeit wird in diesen drei Modellen als Pflichtbewusstsein innerhalb der Arbeitssituation aufgefasst.
[113] Die Ergebnisse werden in dieser Arbeit nicht berichtet.

b) Personenbezogene Überzeugungen im Kontext des beruflichen Beanspruchungserlebens - Substitutionsskalen für das Arbeitsbezogene Verhaltens- und Erlebensmuster nach Schaarschmidt und Fischer

Anders als in den von Klusmann u.a. im Projekt COACTIV durchgeführten Studien wurden die personenbezogen Überzeugungen im Kontext des beruflichen Beanspruchungserlebens nicht mit dem AVEM von Schaarschmidt und Fischer gemessen. Stattdessen wurden Skalen von verschiedenen Forschergruppen kombiniert. Das Arbeitsengagement wurde mit der neun Items umfassenden Kurzversion der Utrechter Work Engagement Scale (UWES) von Schaufeli und Bakker erhoben. Die UWES kann als einfaktorielle Skala aufgefasst werden, durchgeführte konfirmatorische Faktoranalysen legen aber auch eine dreifaktorielle Skala (Vitalität, Hingabe und Aufnahmefähigkeit) nahe (Schaufeli & Bakker, 2003). Statt der vorgesehenen siebenstufigen Skala wurde auch hier eine sechsstufige Skala verwendet.

Statt der Widerstandsfähigkeit wurde das mit der Distanzierungsfähigkeit verwandte Konstrukt des Overcommitment [OC] nach Siegrist erhoben (Jonge, Linden, Schaufeli, Peter & Siegrist, 2008; Siegrist et al., 2004) wozu eine Skala aus dem Effort-Reward Imbalance eingesetzt wurde. Im Vergleich zum AVEM von Schaarschmidt und Fischer wurden folglich die Teilkomponenten Resignationstendenz, Offensive Problembewältigung und Innere Ruhe ausgeklammert. Dies lässt sich einerseits durch die enge Verbindung des Konstrukt Overcommitment mit dem Job-Demand-Resource-Model begründen, andererseits sind die Folgen von zu ausgeprägtem Overcommitment parallel zu den von Schaarschmidt berichteten Befunden von Risiko A-Typen: (Lehr et al., 2009; Schaarschmidt et al., 1999; Siegrist et al., 2004). Schließlich wurde das berufliche Beanspruchungserleben [BEL] mittels einer Skala aus dem Erfurter Belastungs-Inventar erhoben (Böhm-Kasper, Bos, Jaeckel & Weishaupt, 2000). Die verwendeten Skalen sind insgesamt kürzer und somit wird das Risiko von völliger oder item-weiser Nonresponse reduziert. Außerdem stehen die eingesetzten Skalen jedem zu wissenschaftlichen Zwecken frei zur Verfügung. Als Nachteil steht dem eine größere Ungenauigkeit bzgl. der gemessenen Konstrukte gegenüber.

Die GW-Skala wurde als Gewissenhaftigkeit aus dem NEO Big-Five-Inventar nach Costa und McCrae (1992) erhoben, wobei die Items zuerst von Susann Schuster neu übersetzt und mit der deutschen Übersetzung der Big Five-Skala aus dem International Personality Item Pool von Streib und Wiedmaier abgeglichen wurden (Kussau & Brüsemeister, 2007). Das Item „Ich überprüfe meine Arbeit nicht“ wurde durch „Ich überprüfe meine Arbeit“ ersetzt, um Missverständnisse zu vermeiden, die in Fragebögen häufig entstehen, wenn Verneinungen benutzt werden. Ein Vergleich der Items aus den Skalen GW und UWES zeigt eine hohe inhaltliche Übereinstimmung. Daher wurde die GW-Skala zus. auch als Teil der personenbezogenen Überzeugungen im Kontext des beruflichen Beanspruchungserlebens in die entsprechenden Modelle integriert.

c) Das fachdidaktische Selbstkonzept und die fachdidaktische Selbstwirksamkeitserwartung

Das fachdidaktische Selbstkonzept und die fachdidaktische Selbstwirksamkeitserwartung im Mathematikunterricht wurden mit Skalen gemessen, die erst für die Studien dieses Projekts entwickelt wurden. Anders als beispielsweise in den Projekten COACTIV (Baumert et al., 2008) und MT21 (Blömeke & Kaiser et al., 2008) wurde ein fachspezifisches und auf fachdidaktisch-unterrichtliche Anforderungen zielendes Verständnis zugrunde gelegt. Vorher konnte auf fach- und berufsunspezifische Skalen zurückgegriffen werden.[114] Den Studien dieses Projekts ist die Annahme gemeinsam, dass in VERA8 zurückgemeldete Ergebnisse vorwiegend als Bewertung der fachdidaktischen Aufgabenbewältigung der Lehrkräfte zu verstehen sind. Zumindest für das Fähigkeitsselbstkonzept kann zusätzlich eine Substitution des mathematischen Fähigkeitsselbstkonzepts unterstellt werden, da die fachdidaktische Kompetenz (zumindest für Gymnasiallehrkräfte) in einem engen Verhältnis zur fachwissenschaftlichen Kompetenz von Mathematiklehrkräften steht (Blum et al., 2008).

Beide in den Studien A und B eingesetzten Skalen setzen sich aus Items zusammen, die auf der Vorstellung von gutem Unterricht beruhen. Dabei wurden Vorstellungen von Physik-Referendaren, die Merzyn zu ihrer Vorstellung von gutem Unterricht befragt hat (Merzyn, 2006), auf die konkreten Anforderungen des Mathematikunterrichts übertragen und durch Merkmale von Unterrichtsqualität aus wissenschaftlicher Sicht ergänzt, wie man sie bei Helmke findet (Helmke, 2009). Die Itemformulierung geschah für die fachdidaktisch-unterrichtliche Fähigkeitsselbstkonzeptskala analog zu den Formulierungen des SESSKO zur Erfassung des schulischen Selbstkonzepts (Schöne et al., 2002), die Skala zur fachdidaktisch-unterrichtlichen Selbstwirksamkeitserwartung war an die Skala von Schwarzer und Schmitz angelehnt. Die ursprünglich in beiden Skalen zwanzig Items wurden auf sechs (Fähigkeitsselbstkonzeptskala) und acht (Selbstwirksamkeitsskala) kriteriale Items durch Pretests und anschließende Faktorenanalysen reduziert. Außerdem wurden sie als sechsstufige Skala verwendet.

[114] Für das Fähigkeitsselbstkonzept in Mathematik existiert eine von H. W. Marsh für die Allgemeinheit entwickelte Skala (SDQIII, 1992). Während diese Skala im Durchschnitt über alle Bevölkerungsgruppen ausgezeichnete Werte aufweist, scheint sie eher ungeeignet, um das mathematisch Fähigkeitsselbstkonzept von Personen zu erfassen, deren berufliche Tätigkeit große mathematische Kompetenz voraussetzt und bei denen man überproportional große mathematische Kompetenz erwartet. Untersuchungen im Rahmen von COACTIV-R scheinen diese Vermutung für eine Stichprobe aus Mathematik-Referendaren zu bestätigen. Auf einer vierstufigen Skala lag der Mittelwert über alle vier eingesetzten Items bei 3,4 (Varianz 0.53, pers. Auskunft Mareike Kunter am 07.09.2009). Die Selbstwirksamkeitserwartung wird in deutschen Studien vorwiegend mit der von Schwarzer und Schmitz entwickelten Skala zur Messung der Lehrer-Selbstwirksamkeit erhoben (Schmitz, 2000, Schulte 2008). Diese berufsspezifische Skala bildet allgemeine pädagogische Herausforderungen des Schulalltags ab, nicht aber spezifische Herausforderung des Mathematikunterrichts.

6.2.3 Skalen zur Nutzung von Feedbackinformationen in den Studien A und B

In Studie A wurde das Verhalten im Zusammenhang mit der Nutzung von Feedbackinformationen in allgemeiner Form erfasst. Die dabei relevanten Variablen waren mit (c) der Selbstwirksamkeitserwartung [SWE] und (b) der Gewissenhaftigkeit [GW] teilweise deckungsgleich mit den Variablen des Lehrer-Handlungskompetenzmodells. Es mussten aber (d) die Zielsetzungen [ZIELE] sowie (e) die Überzeugungen zur Leistungsattribution [LA] in diesem Modell ergänzt werden.

Studie B stellte die Grundlage für ein Modell dar, welches das Vorbereitungsverhalten in einen direkten Kontext zu VERA8 setzte. Dieses Modell wurde auch nach der Theorie von Kluger und DeNisi über die Nutzung von Feedbackinformationen konzipiert, umfasste aber neben der (fachdidaktischen) Selbstwirksamkeitserwartung, der Gewissenhaftigkeit, (f) den Kernlehrplänen als Indikator für die Akzeptanz der Zielsetzungen auch (g) eine Skala zu externer Unterstützung und (h) drei Skalen zu Rezeption, Reflexion und Unterrichtsveränderungen in Anlehnung an das Rahmenmodell der pädagogischen Nutzung von Vergleichsarbeiten von Helmke und Hosenfeld (2005).

(d) Fachliche Unterrichtsziele im Mathematikunterricht

Der Fragebogen beinhaltete 28 fachliche Unterrichtsziele. Dabei handelte es sich um vier Bereiche fachlich-umfassender Unterrichtsziele aus MT21 („Routineaufbau“, „Argumentieren und Kommunizieren“[115], „Beweisen“ und „affektiv-motivationale Lernziele“ Müller et al., 2008) sowie Unterrichtsziele zu drei der vier Prozesskompetenzen („Modellieren“, „Problemlösen“ und „Werkzeuge nutzen“) aus den Kernlehrplänen NRW. Bei der Auswahl der Unterrichtsziele wurden Bereiche ausgewählt, bei denen eine möglichst breite Streuung zwischen den Lehrkräften zu erwarten war. Gleichzeitig sollten die Unterrichtsziele allen Lehrkräften als mögliche Unterrichtsziele bekannt sein. Es wurden daher keine inhaltlichen Bereiche in die Liste der Unterrichtsziele aufgenommen, da von diesen zu erwarten war, dass alle Lehrkräfte ihnen ein hohes Gewicht beimessen. Dies konnte hingegen für die Unterrichtsziele der Prozesskompetenzen aus den Kernlehrplänen nicht vorausgesetzt werden, da diese als eigenständige Bereiche dort erstmals explizit in nordrhein-westfälischen Lehrplänen benannt werden. Die Bereiche, die dem Projekt MT21 entnommen wurden, decken (mit Ausnahme des Bereichs „Argumentieren und Kommunizieren“) traditionelle Ziele des Mathematikunterrichts ab, können aber als umstritten gelten. Zusätzlich gilt der Bereich „Beweisen“ als besonders herausfordernd.

Die Bereiche „Argumentieren und Kommunizieren“ und „Beweisen“ wurden jeweils mit fünf Items, die Bereiche „Routineaufbau“ und „affektiv-motivationale Lernziele“ wurden mit vier

[115] Im Original heißt die Skala „Argumentieren und Begründen“. Die Items umfassen aber auch Kommunikationsarten, die keine explizite Argumentation oder Begründung beinhalten.

Items abgebildet. In die Analyse wurden aber nur die fünf Items aus dem Bereich „Argumentieren und Begründen“ einbezogen („lernen einen mathematischen Essay zu schreiben“, „lernen, mathematische Ideen zu erläutern bzw. in Worte zu fassen“, „in der Lage sind, über mathematische Inhalte zu kommunizieren“, „die mathematische Terminologie korrekt benutzen können“ und „lernen, mathematisch zu argumentieren“). Die Prozesskompetenzen „Modellieren/Problemlösen“ und „Werkzeuge nutzen“ wurden mit sechs bzw. vier Zielen dargestellt. Damit wurden „Modellieren“ („mathematische Lösungen am Realmodell überprüfen und ggf. Anpassungen am Modell oder der Lösung vornehmen“, „zu mathematischen Modellen passende Realsituationen finden können“) und „Problemlösen“ („lernen, auch neue mathematische Probleme zu lösen“, „selbst Problemlösungen entwickeln“ und „Problemlösestrategien vergleichen und bewerten können“) abweichend von der Einteilung der Kernlehrpläne nicht als getrennte Konstrukte behandelt, sondern entsprechend der Einteilung bei MT21 als gemeinsame Kompetenz. Der Bereich „Argumentieren/Kommunizieren“ wurde nicht erfasst, da eine inhaltliche Übereinstimmung mit dem Bereich „Argumentieren und Begründen“ gegeben ist. Der Bereich „Werkzeuge nutzen“ wurde hingen mit vier Items erfasst („ggf. geeignete Nachschlagewerke, Zeitschriften oder auch Internetsuchmaschinen nutzen“ „die geeignete Möglichkeit kennen und nutzen, um mathematische Lösungen zu präsentieren“, „Hilfsmittel wie Taschenrechner oder Tabellenkalkulationsprogramme sinnvoll anwenden können“ und „wissen, wann ein Taschenrechner benutzt werden soll und wann nicht“).

Die Lehrkräfte wurden gebeten, eine Gewichtung der Unterrichtsziele für ihren Unterricht anzugeben. Dabei waren die Items bipolar mit den Randlabels „lehne völlig ab“ und „stimme völlig zu“ versehen.

e) Überzeugungen zur Leistungsattribution

Die Leistungsattribution und die Bedeutung von Leistungen in der Schule wurden einerseits mit sechs Items von Hartmut Ditton gemessen (Ditton, 2000) und andererseits mit vier selbst entwickelten Items. Für die Analyse wurde schließlich aber nur eine Skala berücksichtigt, die aus vier Items bestand. Dabei handelt es sich um die Items „Misserfolge von Schülern werden an unserer Schule von den Lehrkräften verantwortet“ und „Schulversagen wird bei uns als Problem der Schule und weniger des Schülers gesehen“ aus der ursprünglichen Skala von Ditton u.a. sowie um „Für die Erfolge meiner Schüler trage ich eine große Verantwortung“ und „Misserfolge liegen vor allem in der Verantwortung der Schüler selbst“. Die Bedeutung von Leistung wurde somit nicht berücksichtigt, da sie indirekt eine Variante der Zielsetzung abbildet. Die Verbindung zwischen VERA8 und dieser Form der Zielsetzungen lässt sich nicht deutlich genug nachvollziehen. Die Endpunkte der Antwortmöglichkeiten waren mit „trifft gar nicht zu“ und „trifft völlig zu“ überschrieben.

(f) Die Akzeptanz der Kernlehrpläne

Die sechsstufige Skala zur Akzeptanz der Kernlehrpläne [KL] bestand aus vier Items und deckte die Aspekte Sinnhaftigkeit, Nützlichkeit und Wirkung ab. Für die Skala wurden ebenfalls aus den Interviews der qualitativen Vorstudie acht Items gewonnen, die nach einem Pretest und einer Faktorenanalyse auf vier reduziert wurden. Im Feedback-Nutzungs-Modell M21, das aus der FIT von Kluger und DeNisi abgeleitet wurde, bildete die Skala die Akzeptanz der Ziele ab. Vorausgesetzt wurde hier, dass Lehrkräfte die Kernlehrpläne als ursprüngliche Grundlage von VERA8 in Nordrhein-Westfalen erkennen. Die Skala war bipolar und erneut mit den Randlabels „lehne völlig ab“ und „stimme völlig zu“ konzipiert.

(g) Die erlebte Unterstützung durch die Schulleitung

Die erlebte Unterstützung durch die Schulleitung [USL] war ursprünglich eine sieben Items umfassende unipolare Skala aus Eigenentwicklung. Die Items lassen sich in zwei Teilbereiche trennen, nämlich in Items, in denen das Interesse der Schulleitung an einem guten Abschneiden thematisiert werden, und in solche, die tatsächliche Handlungsreaktionen der Schulleitung auf die Ergebnisse von VERA8 abbilden. In die Analyse wurde aber nur eine Skala einbezogen, die sich aus den vier Items zu den tatsächlichen Handlungsreaktionen zusammensetzte („Wenn eine Klasse in einem Fach bei den LSE… unterdurchschnittlich abschneidet, fragt die Schulleitung die unterrichtende Fachlehrkraft, welche Erklärung sie für das schlechte Abschneiden hat“, …gut abschneidet, hebt die Schulleitung die unterrichtende Fachlehrkraft besonders hervor“, „Wenn die Klassen einer Lehrkraft mehrfach schlechtere Ergebnisse bei den LSE erreichen als andere Klassen an unserer Schule… bittet die Schulleitung die Lehrkraft zum Gespräch“ und „…bietet die Schulleitung Unterstützung an“). Dadurch sollte verhindert werden, Angaben zu berücksichtigen, über die die Befragten möglicherweise nur spekulieren konnten.

(h) Das Interesse an VERA8-Ergebnissen, der Umgang mit Ergebnissen und vorgenommene Veränderungen

Die drei Skalen zum Interesse an VERA8-Ergebnissen [REPZ], dem Umgang mit vorherigen Ergebnissen [RFL] und aufgrund von VERA8-Ergebnissen herbeigeführten Veränderungen [VEA] decken die drei Stufen Rezeption, Reflexion und (Re-)Aktion aus dem Modell der Unterrichtsentwicklung nach Helmke und Hosenfeld (2005) ab und wurden aus der Befragungen im Rahmen von VERA3 übernommen (Groß Ophoff, 2013). Dabei wurden für die Messung der Veränderungsbereitschaft nur die ersten vier Items einbezogen, die als „individuelle Veränderungen“ überschrieben waren. Die vier Items zu „Veränderungen durch die Fachgruppe“ wurden nicht berücksichtigt. Die beiden anderen Skalen wurden jeweils vollständig übernommen und umfassen je fünf Items. Sie waren im Fragebogen mit unipolarer Antwortauswahl dargestellt.

Tabelle 6.2

Skalenübersicht zu den Items des jeweils zweiten Teils der Fragebögen

Name	Abkürzung	Itemanzahl	Fragebogen (Frage)	Modelle	Quelle
Transmitterüberzeugungen	TÜ	5	1 (15)	M12 & M12a	(Blömeke & Müller et al., 2008)
Konstruktionsüberzeugungen	KÜ	6	1 (15)	M12	(Blömeke & Müller et al., 2008)
Utrechter Work Engagement Scale	UWES	9	1 (18)	M11, M11a, M12 & M12a	(Schaufeli & Bakker, 2003)
Overcommitment	OC	6	1 (20)	M11, M11a, M12 & M12a	(Siegrist et al., 2004)
berufliches Beanspruchungserleben	BEL	7	1 (19)	M11, M11a, M12 & M12a	(Böhm-Kasper et al., 2000)
Gewissenhaftigkeit	GW	10	1 (21) & 2 (20)	M11a, M21 & M22	(Kussau & Brüsemeister, 2007)
fachdidaktisches Selbstkonzept	SK	6	1 (16) & 2 (17)	M11, M11a, M12 & M12a	Eigenentwicklung
fachdidaktische Selbstwirksamkeitserwartung	SWE	8	1 (17) & 2 (18)	alle Modelle	Eigenentwicklung
Argumentieren/ Begründen	ZieleArK	5	1 (22)	M21	(Müller et al., 2008)
Problemlösen/ Modellieren	ZielePLM	6	1 (22)	M21	(Müller et al., 2008), Eigenentwicklung
Werkzeuge nutzen	ZieleW	4	1 (22)	M21	Eigenentwicklung
Leistungsattribution	LA	4	1 (14)	M21	(Ditton, 2000), Eigenentwicklung

Name	Abkürzung	Itemanzahl	Fragebogen (Frage)	Modelle	Quelle
Akzeptanz der Kernlehrpläne	KL	4	2 (14)	M22	Eigenentwicklung
Unterstützung durch die Schulleitung	USL	4	2 (23)	M22	Eigenentwicklung
Interesse an VERA8-Ergebnissen	REPZ	5	2 (21)	M22	(Groß Ophoff, 2013)
Umgang mit vorherigen Ergebnissen aus VERA8	RFL	5	2 (25)	M22	(Groß Ophoff, 2013)
herbeigeführte Veränderungen	VEA	4	2 (28)	M22	(Groß Ophoff, 2013)

6.3 Reflexion der Datenauswertung

Im letzten Abschnitt dieses Kapitels über die methodische Anlage der dieser Dissertation zugrunde liegenden Studien wird ein kurzer Blick auf die durchgeführten Auswertungen geworfen. Dabei wird zuerst die Auswertung des jeweils ersten Teils der beiden Fragebogenversionen begründet, die sich mit der Qualität und dem Umfang der Vorbereitung auf VERA8 befasst. Zweitens wird erläutert, warum und in welcher Weise latente Klassenanalysen [LCA] durchgeführt wurden, um Experten-Modelle zu vergleichen und auch die Skalenbildung mittels LCA vorgenommen wurde.

6.3.1 Deskriptive Auswertung des Testcoachings

Bei den jeweils im ersten Teil der beiden Fragebogenversionen erhobenen Daten handelt es sich überwiegend um nominale oder ordinale Daten. Eine Ausnahme stellt Frage 1 („Wie viele Unterrichtsstunden haben Sie in dieser Klasse ungefähr für die unmittelbare Vorbereitung auf die LSE aufgewendet?“) dar, mit der tatsächliche metrische Daten gewonnen wurden. Für die metrischen Daten bzw. die als metrisch angenommenen Daten werden in der Auswertung jeweils die Anzahl der berücksichtigten Fälle [n], das arithmetisches Mittel [M], die Standardabweichung [SD] und dazu ein Konfidenzintervall [95% CI] mit zwei Nachkommastellen angegeben, welches auf Basis von 1000 Bootstrapstichproben berechnet wurde.

Für Frage 1 (Anzahl der Vorbereitungsstunden) wurden verschiedene Gruppenvergleiche durchgeführt. Dabei kamen gewöhnliche statistische Test zum Einsatz, um die jeweilige konkrete Ausformulierung der Hypothese H_0: „Die beiden Gruppen unterscheiden sich bzgl. des getesteten Merkmals nicht voneinander“ zu überprüfen. Sofern die Daten in den zu vergleichenden Stichproben als normalverteilt angenommen werden konnten und Varianzhomogenität bestand, handelte es sich dabei um den t-Test nach Student. Die Daten wurden für die (Teil-)Stichproben jeweils mit dem Kolmogorov-Smirnov-Test zu Niveau $\alpha=0.01$ bzgl. einer Normalverteilung und mit Levene-Test zu $\alpha=0.05$ bzgl. der Varianzhomogenität getestet. Statt eines t-Test nach Student soll im Fall nicht normalverteilter Daten ein parameterfreier Test verwendet werden (Bortz, Lienert & Boehnke (2008)). Der Wilcoxon-Rangsummen-Test bzw. der äquivalente Mann-Whitney-U-Test hatten sich als parameterfreier Test angeboten, setzen aber ebenfalls Varianzhomogenität voraus. Im Fall von signifikant unterschiedlicher Varianz wurde auf den verallgemeinerten t-Test nach Welch zurückgegriffen. Dies geschah auch im Fall zusätzlich fehlender Normalverteilung der Daten. Bortz, Lienert und Boehnke (2008) empfehlen bereits ab einem Stichprobenumfang von $n>30$ mit Verweis auf den zentralen Grenzwertsatz eine Normalverteilung zu unterstellen und parametrische Tests anzuwenden. Das grundsätzlich

angenommene Niveau für den Fehler erster Art betrug jeweils α=.05. Nur im Fall der Ablehnung von H_0 wurde auch mit α=.01 getestet und wird entsprechend angegeben. Berichtet werden dem Standard folgend der ermittelte t-Wert [t], die Freiheitsgrade [df] und der P-Wert [p].

Bei signifikanten Mittelwertunterschieden wurde die Effektstärke [r] als Effektstärke nach Cohen für Mittelwertunterschiede als Differenz der Mittelwerte normiert durch die (allerdings gewichtete) gemeinsame Varianz berechnet und anschließend normiert. Für die Einordnung der Werte fehlt es an spezifischen Referenzgrößen. Als Faustregel gilt aber, dass als kleine Effekte Effekte mit r>.10, mittlere Effekte mit r>.234 und große Effekte mit r>.371 angesehen werden können (Volker, 2006).

Auch für die Normal- und Ordinalskalen wurden Gruppenvergleiche durchgeführt. Dazu werden die absoluten Verteilungen sowie die relativen Häufigkeiten ausgewiesen. Sofern nur ein absoluter und relativer Wert angegeben ist, handelt es sich um dichotome Skalen. Als statistischer Test musste auf einen voraussetzungsarmen χ^2-Homogenitätstest zurückgegriffen werden (Weick, 1976). Berichtete werden der empirisch ermittelte χ^2-Wert, die Anzahl der Freiheitsgrade sowie der P-Wert. Die Effektstärke für Nominalskalen wird mit Cohens w angegeben (Volker, 2006).

6.3.2 Skalenbildung und die Vergleiche der Experten-Modelle mittels latenter Klassenanalysen

Sowohl für die Vergleiche zwischen den verschiedenen Experten-Modellen als auch für die Skalenbildung wurden latente Klassenanalysen [LCA] durchgeführt. LCA dienen in erster Linie dazu, Personengruppen aufgrund von qualitativen Merkmalen zu unterscheiden. Die Typisierung folgt dabei der Vorstellung, dass nur die qualitativen Merkmale als kategoriale Variablen erfassbar sind, die kategoriale Gruppenvariable hingegen latent ist. Als grundlegende Parameter müssen für jede Ausprägung einer jeden Variablen für alle angenommenen Gruppen bedingte Wahrscheinlichkeiten berechnet werden. Unter der Festhaltung von einer Klasse, einer Variablen und einer ihrer spezifischen Ausprägungen ist die Lösungswahrscheinlichkeit (gleichbedeutend mit dem Auftreten dieser Ausprägung) von der Person unabhängig konstant (Rost, 2004).

Die Klasseneinteilung erfolgt nach Algorithmen, die ähnliche Antwortmuster (Personen) in dieselbe Klasse einordnen, sodass die Interklassenheterogenität und die Intraklassenhomogenität maximiert werden. Die Beziehung von Klassenzuordnung, Variablen und deren Ausprägungen auf der einen sowie der Antwortmuster aus dem Datensatz auf der anderen Seite drückt sich in der Likelihoodfunktion [L] aus. Es gilt dasjenige Klassenmodell als das optimale, welches den Wert L maximiert. Dazu nutzt man Algorithmen, die Spezialfälle des Expectation-Maximization-Algorithmus [EM-Algorithmus]

darstellen (z.B. nach Goodman vgl. Kussau, 2007). Diese schätzen die unbekannten Parameter und führen dann einen Maximum-Likelihood-Ansatz durch. Numerische Probleme außer Acht gelassen konvergieren diese Verfahren in jedem Fall, allerdings ggf. nur gegen ein lokales Maximum, sodass das Finden des globalen Maximums evtl. mehrere Durchläufe benötigt. Das typische Vorgehen ist daher, die LCA mit meist fünfzig Durchläufen und gleichbleibenden vorgegebenen Modellparametern durchzuführen und anschließend die Modellwerte zu nutzen, bei denen die Likelihood-Funktion den größten Wert angenommen hat. Eine Garantie, das tatsächliche globale Maximum gefunden zu haben, besitzt man damit allerdings noch nicht. Andersherum stellen die beobachteten Antwortmuster die zur Bestimmung der unbekannten Modellparameter verfügbaren Gleichungen dar. Für eine eindeutige Identifikation der zu schätzenden Werte müssen folglich mehr Antwortmuster möglich sein als unabhängige Modellparameter zu schätzen sind. Die Zahl der möglichen Antwortmuster ergibt sich für **k Items** mit **m** potenziellen **Ausprägungen** als

$$N_t = m^k.$$

Klasse an (disjunkte und exhaustive Klassen). (2) Alle Items messen mit der Klassenzuordnung dieselbe Personenvariable (Itemhomogenität). (3) Innerhalb einer Klasse besteht zwischen den Items eine lokale stochastische Unabhängigkeit (Rost, 2004)[116].

Der Umgang mit den drei Voraussetzungen ist allerdings nicht ganz konfliktfrei. LCA-Modelle gehören zu den Modellen der Item-Response-Theorie, sodass die Zuweisung zu einer Klasse nicht deterministisch, sondern probahilistisch erfolgt (Gollwitzer, Mossbrugger & Kelave, 2008). Gleichzeitig erschwert die probahilistische Zuweisung die Interpretation und weitere Nutzung der Klassenzuweisung. Möglich wäre es, die Zuweisungsvektoren ähnlich zum Umgang mit den Zuweisungen unter kontinuierlichen Modellen der Item-Response-Theorie als (diskrete) Verteilungsfunktion zu modellieren und jedem Antwortmuster entsprechend der Zuordnungswahrscheinlichkeit zu den Klassen Plausible Values zuzuordnen. Weiterführende Analysen müssten dann jeweils entsprechend der Anzahl dieser Plausible Values durchgeführt werden. Üblich ist stattdessen die Verwendung des Modalwerts, da dieses Vorgehen den Klassifikationsfehler minimiert.

Die Zahl der zu schätzenden Modellparameter lautet bei *G* Klassen und *k m*-stufigen Items für das nicht-restringierte Modell:

$$n_p = G \cdot k \cdot (m - 1) + G - 1$$[117]

[116] Bei Rost sind dies die Voraussetzungen (2)-(4), die bei Rost mit (1) gekennzeichnete Annahme wurde bereits im ersten Absatz beschrieben.

[117] Im Unterschied zu der von Rost (2004) gewählten Darstellung steht m für die Anzahl der Antwortmöglichkeiten, nicht für die Anzahl der Schwellenparameter.

Der linke Summand steht dabei für die *m-1* Antwortwahrscheinlichkeiten, welche pro Variable *k* und pro Gruppe *G* berechnet werden müssen und für jede Variable in jeder Gruppe jeweils in Summe eins ergeben müssen. Der rechte Summand steht für die *(G-1)* zu berechnende Klassenhäufigkeit.

Die LCA gehört zu den strukturgebenden Verfahren. Statt einer LCA könnten folglich auch Clusteranalysen durchgeführt werden. Anders als bei Clusteranalysen, die ebenfalls Personen in Gruppen einteilen, wird bei der LCA aber kein Ähnlichkeits- oder Distanzmaß festgelegt und keine Distanz zwischen den Clustern definiert. Dies ermöglicht die Einteilung mittels LCA bereits für kategoriale Items, während Clusteranalysen Daten voraussetzen, auf denen eine Metrik definiert werden kann. Schließlich besitzt die LCA den Vorzug, die Anzahl der Klassen statistisch prüfen zu können (Kussau, 2007). Hierin spiegelt sich der explorative Charakter der LCA wider. Bei einer LCA muss die Anzahl der Klassen vorgegeben werden, während die Klassengröße und die Zuweisung der Antwortmuster (Personen) zu den einzelnen Klassen ein Ergebnis des Algorithmus sind. Durch Durchläufe des Algorithmus mit verschiedenen Klassenanzahlen und den ermittelten Parametern lässt sich dann die Klassenlösung auswählen, die am besten zu den Daten passt.

Der Vergleich der verschiedenen Klassenlösungen mit unterschiedlicher Anzahl an Klassen erfolgt in erster Linie über Informationskriterien. Dabei wird der Wert der maximierten Likelihood-Funktion in Relation zu der Anzahl der Modellparameter gesetzt. Maßgeblich ist die Idee, dass ein Modell mit vielen Parametern automatisch den Datensatz besser erklärt, viele Modellparameter aber dem Gebot der Einfachheit gegenüberstehen. Genutzt werden können das Akaike Infomation Criteroin

$$AIC = -2\log(L) + 2n_p,$$

das Bayesian Infomation Criteroin

$$BIC = -2\log(L) + (\log(N)) \cdot n_p$$

und das Consistent AIC

$$CAIC = -2\log(L) + (\log(N)) \cdot n_p + n_p.$$

Die letzten beiden Kriterien berücksichtigen neben den zu schätzenden Modellparametern zusätzlich die Stichprobengröße. Das wurde das CAIC besonders für größere Stichproben entwickelt. Rost (2004) schlägt vor, vorwiegend dem AIC Gewicht zu verleihen, wenn die Itemzahl klein und die Zahl der Antwortmuster groß ist, während das BIC betrachtet werden soll, wenn die Itemzahl groß und kleinere Patternhäufigkeiten vorliegen. Für alle drei Maße gelten Modelle mit den kleineren Werten als die besseren Modelle.

Die drei Kriterien lassen sich aber nur für einen Vergleich verschiedener Modelle nutzen, um die Klassenzahl zu bestimmen. Sie lassen keine Aussage darüber zu, wie gut das Modell zum Datensatz passt. Dazu müssen globale Maße genutzt werden, die den Wert L in Beziehung zu einer χ^2-Statistik setzen. Prinzipiell sind zwei Arten von Tests möglich: Der Likelihood-Quotienten-Test [LQT] setzt die L-Werte zweier Modelle in ein Verhältnis. Wenn L_0 ein Obermodell ist von L_1, d.h. wenn L_1 durch Restriktionen aus L_0 gewonnen wird, dann gilt

$$\chi^2 \leftarrow -2 \log(LR) \text{ mit } LR = L_1/L_0 \text{ und } df = n_p(L_1)/n_p(L_0).$$

Bedingung ist aber, dass L_1 nicht durch Nullsetzen von zu schätzenden Parametern aus L_0 entstanden ist, d.h. die optimale Klassenzahl lässt sich so nicht ermitteln. L_0 muss außerdem ein bereits gültiges Modell sein. Normalerweise besitzt man als Vergleichsmodell nur das saturierte Modell L_{sat}, welches die beobachteten Daten deswegen perfekt erklärt, weil es genauso viele Parameter besitzt wie es unabhängige Daten im Datensatz gibt. Überprüft wird mit dem LQT, ob das Modell L_1 die Datenstruktur genauso gut erklärt wie L_{sat}. Getestet wird auf Beibehaltung der Nullhypothese, sodass der LQT-Wert nicht größer als der χ^2-Wert zur entsprechenden Anzahl an Freiheitsgraden df und dem gewählten Niveau α sein darf.

Letzteres gilt auch für die zweite Art der globalen Modelltests, den Pearson'schen χ^2-Test CHI. Hierbei werden die Häufigkeiten der beobachteten Datenmuster n_o mit den aufgrund des berechneten Modells zu erwartenden Häufigkeiten n_e verglichen:

$$\text{CHI} = \text{Summe über alle Antwortmuster } (n_o - n_e)^2/n_e \text{ mit } df = m^k - n_p - 1$$

Beide Ansätze setzen allerdings voraus, dass die Zahl n_e für alle Pattern größer als eins sein muss. Anderenfalls folgt die Prüfgröße nicht mehr der χ^2-Verteilung (Rost, 2004). Read und Cressie (1988) haben daher als Ersatz für den Pearson'schen χ^2-Test CHI einen verallgemeinerten χ^2-Test entwickelt. Dieser benötigt allerdings die Spezifikation einer Konstanten λ. Sie schlagen für die LCA den Wert $\lambda = ⅔$ vor (Jensen & Meckling, 1976). Bisher hat sich dieser Test allerdings nicht etabliert (Ross, 1973). Sind zu viele Antwortmuster nicht beobachtet worden oder wurde für sie ein n_e kleiner als eins geschätzt, besteht nur die

Möglichkeit, spezielle Verteilungen mittels parametrischem Bootstrapping zu erzeugen (Davier, 1997). Dabei werden neue Datensätze generiert, auf die das ursprünglich geschätzte Modell passt. Für diese Datensätze berechnet man dann ein äquivalentes Modell und die zugehörigen Prüfgrößen. So erhält man aus den Prüfgrößen eine neue Statistik, die mit der ursprünglichen Prüfgröße verglichen werden kann. Liegen α der Werte aus der Statistik über der ursprünglichen Prüfgröße, fittet das Modell die Daten. Dieses Vorgehen wurde auch für die Modellvergleiche dieser Arbeit genutzt, wobei jeweils 2000 Datensätze generiert wurden.

Die Skalenbildung

In beiden Fragebogenversionen waren die Skalen mit einer sechsstufigen Antwortskala versehen. Für jede Skala ergibt sich dadurch schon ein enormer Umfang an möglichen Antwortmustern. Zusätzlich wären durch die LCA berechnete Experten-Klassen durch die Vielzahl an Items auf Itemebene nicht interpretierbar. Die Skalen mussten folglich einer Datenreduktion unterzogen werden. Ein durchaus gebräuchliches Verfahren für eine solche Datenreduktion ist ein Mediansplit, allerdings sind die Daten dann schwieriger inhaltlich zu interpretieren. Eine andere Möglichkeit kann eine Gruppierung der Probanden am theoretischen Mittel des jeweiligen Summenscores sein (Hosenfeld, 2010). Beide Verfahren setzen aber voraus, dass es sich um eine eindimensionale kontinuierliche Persönlichkeitseigenschaft handelt. Daher bieten sich restringierte Klassenanalysen demgegenüber als Alternative an (Kieser & Ebers, 2006; Rost, 2004).

Die latenten Klassenanalysen wurden, da es sich im Fragebogen jeweils um Itembatterien handelte, als LCA für ordinale Ratingskalen berechnet. Ein wesentlicher Vorteil dieses Modells für ordinale Daten ist, dass für diese Modellklasse weniger unabhängige Parameter geschätzt werden müssen. Nimmt man an, dass die Antworten einer ordinalen Datenstruktur folgen, lautet die Zahl der zu schätzenden Modellparameter bei G Klassen und für das nicht-restringierte Modell:

$$n_p = G \cdot k + k \cdot (m - 2) + G - 1$$

Die zu schätzende Parameterzahl ist damit mindestens um einen Parameter kleiner als die ursprüngliche Zahl, da für jedes der *k* Items mit *m* Antwortmöglichkeiten nur *(m – 2)* Schwellenparameter und *(G · k)* Zustimmungstendenzen zu schätzen sind. Klassenmodelle für Ratingskalen bei ordinalen Daten unterliegen weiter der Annahme, dass die Testpersonen für jedes Item äquivalente Antwortskalen erkennen. Die bei ordinalen latenten Klassenanalysen zu schätzenden Schwellenparameter werden als Eigenschaft des Antwortformats betrachtet, nicht als Teil eines speziellen Items. Die Zahl der letztendlich zu

schätzenden unabhängigen Parameter reduziert sich folglich weiter dadurch, dass für jede Klasse nur noch *(m-2)* Schwellenparameter geschätzt werden müssen statt k-mal so viele im ursprünglichen Modell:

$$n_p = G \cdot k + (m - 2) + G - 1$$

Neben einer reduzierten Zahl an unabhängig zu schätzenden Parametern besitzen speziell Ratingmodelle für die LCA (im Gegensatz zu IRT-Modellen) zusätzlich den Vorzug, dass negativ gepolte Items nicht umgepolt werden müssen (Rost, 2004). Dadurch muss auch nicht unterstellt werden, dass die Testteilnehmenden das Antwortformat als symmetrisch angesehen haben. Problematisch bei der Skalenbildung mittels LCA ist allerdings, dass nicht erkennbar ist, ob die Klassen durch Unterschiede im generellen Antwortverhalten statt durch Iteminhalte bedingt sind (Gollwitzer et al., 2008). Datenmuster, die vollständig aus dem kleinsten oder vollständig aus den größten Ausprägungen bestehen, werden daher nicht in die Schätzungen miteinbezogen (ein Nachteil des JML-Schätzers). Da aber gleichzeitig auch eine Verschiebung dieser Tendenz oder eine „Tendenz zur Mitte“ denkbar ist, wurden extreme Datenmuster nachträglich per Hand doch zugeordnet.

Die Skalenbildung geschah in dieser Arbeit vor allem aber durch die LCA aus der Überlegung heraus, dass es sich bei den gemessenen Persönlichkeitseigenschaften einerseits um latente Variablen handelt und es andererseits für jede der Persönlichkeitseigenschaften mindestens eine Stufe gibt, ab der man als Experte angesehen werden kann. Dem folgend wurden für alle Skalen zuerst Zweiklassenlösungen berechnet und für die weitergehende Analyse der Skalen genutzt. Vergleiche mit anderen Klassenlösungen zeigten aber bessere AIC- Werte (ebenso bei den beiden anderen Informationskriterien) für Modelle mit mehr als zwei Klassen. Daher wurden alle Berechnungen auch auf der Grundlage einer Vierklassenskalierung berechnet.

Die Berechnung wurde mit dem R-Paket mixRasch von Willse durchgeführt. Das Paket nutzt einen Joint Maximum Likelihood Estimation, da es eigentlich für mixed Rasch-Modelle programmiert wurde (Neuweg, 2010). Beim JMLE werden Item- und Personenparameter gleichzeitig geschätzt, der Schätzer ist aber aufgrund der eher geringen Itemzahl weniger konsistent (Rost, 2004). Nutzt man mixRasch allerdings nur für eine LCA, werden die Personenparameter auf null gesetzt und das Problem entfällt. Das Paket kann mit fehlenden Werten als Leerstellen umgehen und wurde mit folgenden Einstellungen genutzt: steps = 5, max.iter = 50, conv.crit = 0.001, model = "RSM", n.c = 2 oder 4, treat.extreme = 0.3, maxchange = 1.5, maxrange = c(-4, 4), as.LCA = TRUE.

Die jeweils berechneten Klassenzuweisungen wurden anschließend derart in ihrer Reihenfolge umkodiert, dass eine ordinal-ähnliche Struktur entsteht, wobei die „beste“ Klasse mit „1“ kodiert wurde. Es wurde aber explizit keine ordinale, sondern lediglich eine

kategoriale Struktur unterstellt. Ergab die Klassenzuteilung keine derart ersichtliche Aufteilung, wurde daher eine inhaltliche Zuweisung vorgenommen. Für die Skala SWE ergab die LCA mit zwei Klassen keine sinnvoll interpretierbare Lösung. Für diese Klasse wurde nur eine Vierklassenlösung berechnet und die Klassen per Hand dichotomisiert, indem jeweils zwei aufgrund der mittleren Zustimmung benachbarte Klassen zusammengefasst wurden.

Die Modellvergleiche

Die LCA gilt in allgemeiner Form als ein exploratives Verfahren, weil sie keine präexperimentellen Hypothesen über die Antwortmuster voraussetzt (Rost, 2004). Trotzdem ist ebenso eine konfirmatorische Nutzung möglich wie die Skalenbildung zeigt. Auch die Modellvergleiche haben ein exploratives wie auch ein konfirmatorisches Element. Explorativ wurde die LCA insofern genutzt, dass grundsätzlich keine Restriktionen vorgenommen wurden. Dies hatte zur Folge, dass bei den Modellvergleichen die latenten Klassen mittels eines anderen R-Pakets berechnet werden mussten, weil mixRasch ordinale Daten annimmt. Zum Einsatz kam das R-Paket poLCA von Linzer und Lewis. Dieses nutzt einen EM-Algorithmus und dabei einen Newton-Raphson-Algorithmus (Lipowsky, 2010). Das Paket berechnet die Informationskriterien AIC und BIC (nicht aber das CAIC) und die globalen Fit-Kriterien des Pearson'schen χ^2-Tests und den LQT, nicht aber den modifizierten χ^2-Test nach Read und Cressie. Letzterer wurde daher auch nicht bei der Entscheidung berücksichtigt, ob die Passung der jeweiligen Modelle auf den Datensatz gegeben ist. Stattdessen wurden für die ersten beiden globalen Fit-Kriterien Statistiken mittels Bootstrapping berechnet. Dazu konnte auf die Möglichkeit zurückgegriffen werden, mit poLCA dem geschätzten Modell entsprechende neue Datensätze zu simulieren, es waren aber auch eigenständige Fortführungen des Programmcodes nötig (s. Anhang). Für die Modelle wurden Statistiken zum Pearson'schen χ^2-Test und den LQT mit jeweils 2000 simulierten Stichproben berechnet. Ausschlaggebend im Vergleich der verschiedenen Modelle war aber neben den Informationskriterien vor allem die Passung anhand des χ^2-Werts, da der LQT bei polynominalen Daten ungenau sein kann (Davier, 1997).

Verglichen wurden die folgenden Modelle[118]:

Das Modell M11 bestand aus den fünf personenbezogenen Überzeugungsskalen (UWES, BEL, OC, SWE & SK) als Prädiktoren für die ermittelte Expertengruppe jeder Lehrkraft. Die Idee hinter dem Modell war die Hypothese, dass das Vorbereitungsverhalten kein fachspezifisches und abgeleitetes Unterrichten ausdrückt. Daher wurden neben dem Wissen & Können auch keine Unterschiede in den Überzeugungen berücksichtigt, welche Ziele im Mathematikunterricht erreicht werden sollten und wie das Lernen optimal funktioniert. Stattdessen wurde angenommen, dass die Vorbereitung vor allem eine Frage dessen ist,

[118] Die in den einzelnen Modellen einbezogenen Ressourcen sind dunkel dargestellt.

welche Ressourcen der Lehrkraft grundsätzlich zur Verfügung stehen. Diese drücken sich sehr deutlich in den personenbezogenen Ressourcen aus. – Das Modell M11a ergänzte die fünf Skalen um die Skala zur Gewissenhaftigkeit [GW]. Dadurch wurde der Bereich Arbeitszufriedenheit und Arbeitsengagement mit einer zweiten Skala ggf. auf eine sichere Basis gestellt. Zusätzlich sollte auch geprüft werden, ob die Nähe zwischen der UWES und der GW, die sich in den Items ausdrückt, tatsächlich auch in diesem Modell wiederfinden lässt.

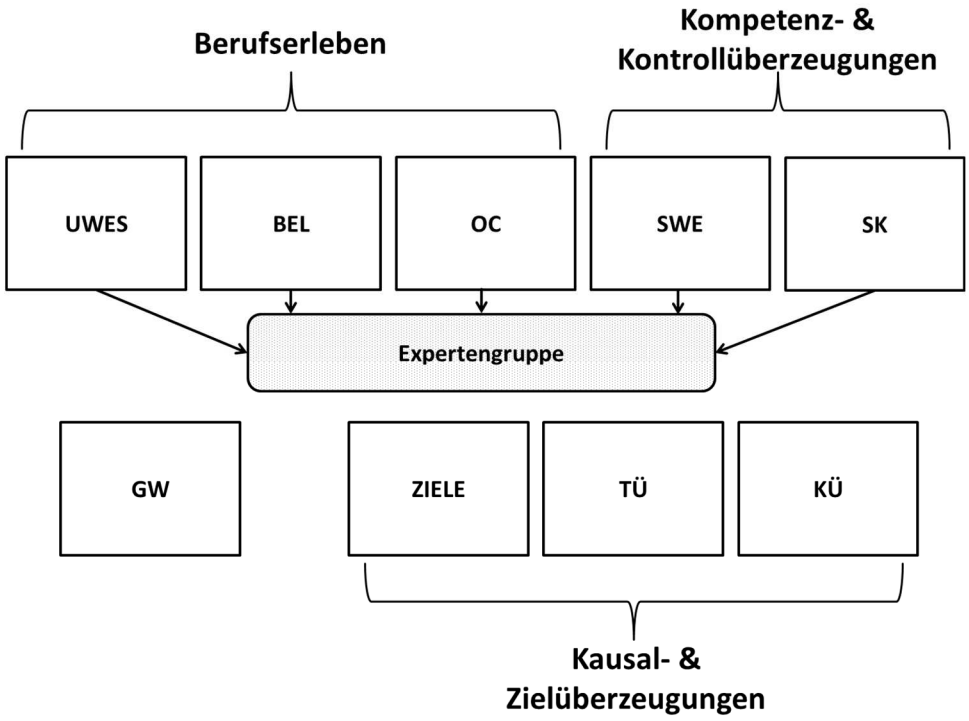


Abbildung 6.1: Modell M11 - personenbezogene Überzeugungen als Prädiktoren des Vorbereitungsverhaltens

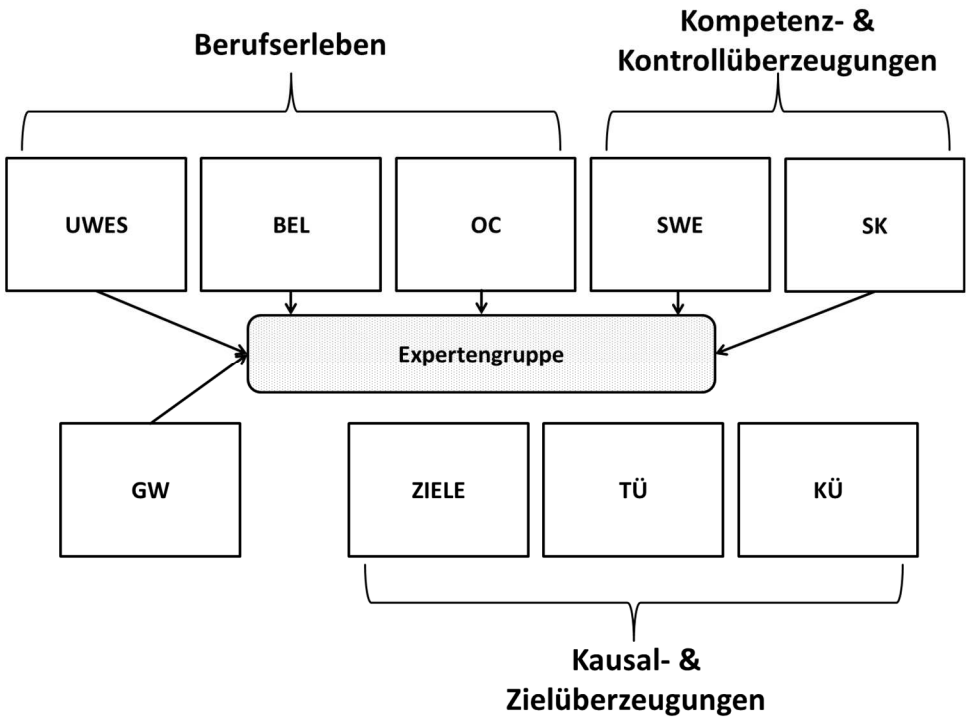


Abbildung 6.2: Modell M11a - personenbezogene Überzeugungen und Gewissenhaftigkeit als Prädiktoren des Vorbereitungsverhaltens

Im Modell M12 wurden auch Ziel- und Kausalüberzeugungen über das Lehren & Lernen als zusätzliche Prädiktoren integriert. Hier wurde hingegen das Vorbereitungsverhalten auch ausdrücklich als Resultat fachspezifischer Planung betrachtet. Die Hypothese zu diesem Modell nahm an, dass Lehrkräfte unterschiedlich vorbereiten, auch abhängig davon, welche Vorstellung sie über das Lehren & Lernen speziell in Mathematik haben. Das Modell ist eine Erweiterung zu Forschungsfrage 2. Aufgrund der punktuellen vorherigen Befunde sollten konstruktions-orientierte Lehrkräfte sich in gleicher Weise verhalten wie Lehrkräfte, die aufgrund der personenbezogenen Überzeugungen eher zu „gutem“ Unterricht fähig sein sollten als andere Lehrkräfte. Spezifische Hypothesen konnten hierzu allerdings nicht formuliert werden, da keine Theorie über eine konkrete Gruppenklassifikation vorlag und die LCA hier explorativ genutzt wurde. Lehr-/Lernziele wurden in diesem Modell nicht berücksichtigt, um das Modell nicht überkomplex anzulegen. Es ging hier in einem ersten Schritt darum zu prüfen, ob überhaupt eine Klasseneinteilung gefunden wird, die sich inhaltlich sinnvoll interpretieren lässt. Aus weitergehenden Analysen von Unterschieden im Vorbereitungsverhalten lassen sich in so einem Fall lediglich erst spezifische Hypothesen formulieren.

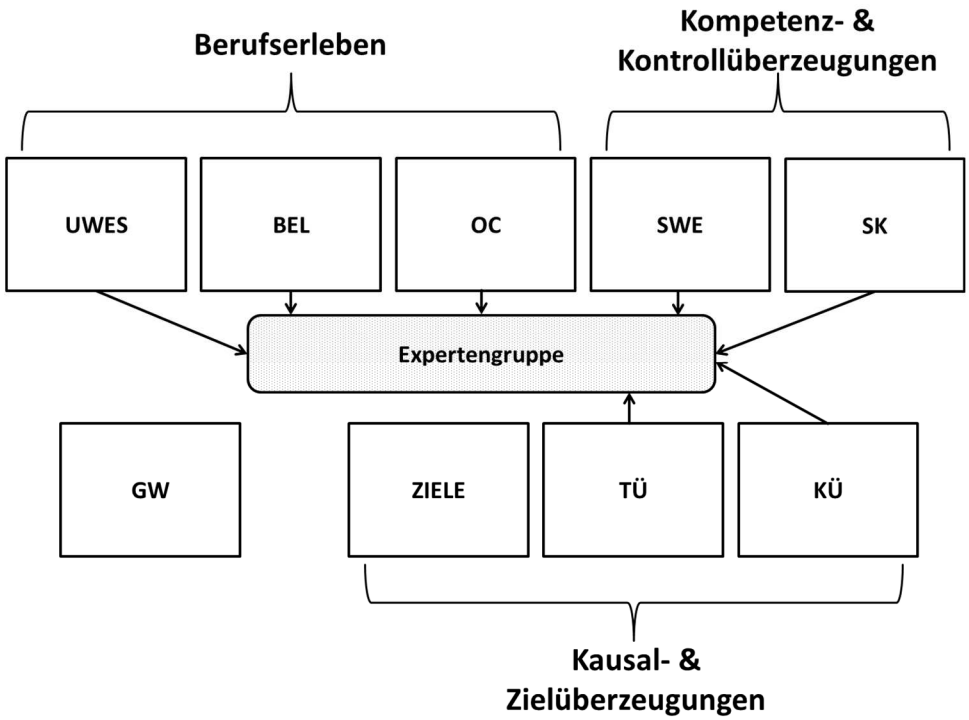


Abbildung 6.3: Modell M12 – personenbezogene Überzeugungen und gegenstandsbezogene Überzeugungen als Prädiktoren des Vorbereitungsverhaltens

Den drei Modellen mit Bezug zur Lehrer-Handlungskompetenz standen zwei Modelle gegenüber, die auf Grundlage der Feedback Intervention-Theorie von Kluger und DeNisi sowie nachfolgenden Forschungsbefunden beruhten.

Das Modell M21 wurde ebenfalls innerhalb der Studie A überprüft. Es verband die fachdidaktische Selbstwirksamkeitserwartung und die Gewissenhaftigkeit mit der Leistungsattributionsüberzeugung (Leistungsattribution [LA] und Bedeutung von Leistung [BL]) und fachspezifischen Lernzielen für den Mathematikunterricht. Von den insgesamt sieben Lernzielen wurden vier bzw. drei ausgewählt: „Argumentieren und Kommunizieren“ [ZIELE ArK], „Problemlösen/Modellieren“ [ZIELE PLM][119] und „Werkzeuge nutzen“ [ZIELE W]. Das Modell nutzte Skalen, die keinen direkten VERA8-Bezug aufweisen. Gleichzeitig gab es dadurch eine Schnittmenge mit dem Modell M11a.

[119] Die Faktorstruktur aller 28 Lernzielitems wurde mittels einer Faktorenanalyse überprüft. Trotz sehr guter Werte im Reliabilitätskoeffizienten ($\alpha > .80$) der eigenständige Skalen bilden „Problemlösen“ und „Modellieren“ hiernach eine gemeinsame Skala. Dieser Befunde zeigte sich bereits schon bei TEDS-M, sodass die beiden Prozesskompetenzen zu einer Zielgruppe zusammengefasst wurden.

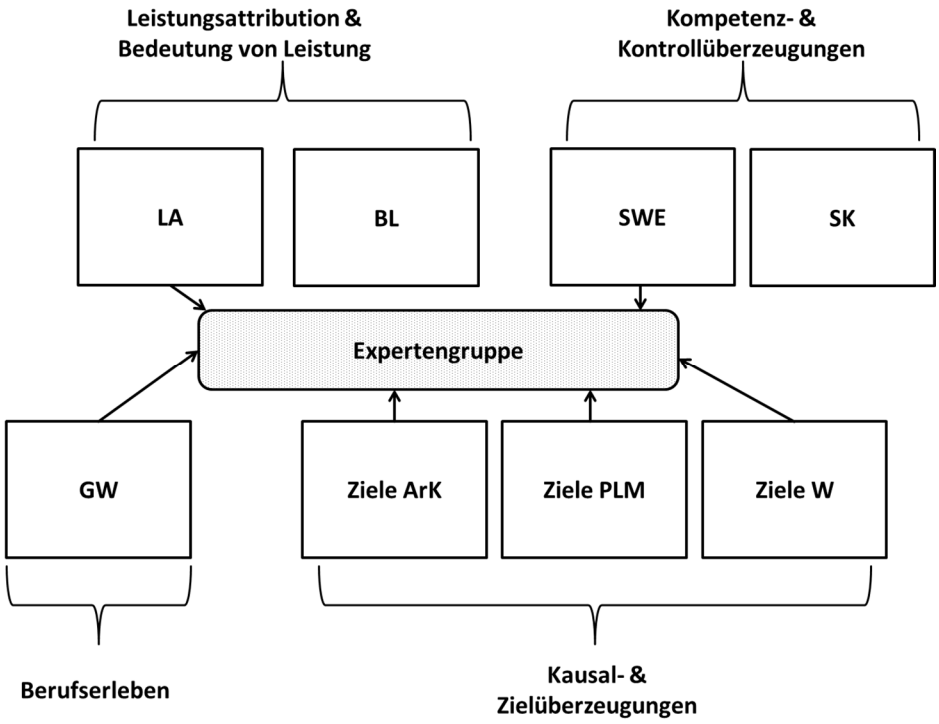


Abbildung 6.4: Modell M21 – allgemein feedback-relevante Überzeugungen als Prädiktoren des Vorbereitungsverhaltens

Das Modell M22 verknüpfte nun die Feedback-Interventions-Theorie und das Zyklusmodell von Helmke und Hosenfeld, indem die Expertengruppen einerseits auf Grundlage der Gewissenhaftigkeit, der fachdidaktischen Selbstwirksamkeitserwartung und der Unterrichtsziele (statt spezifischen Items zu Prozesskompetenzen fanden Items zu den Kernlehrplänen Verwendung) und andererseits auf Grundlage der drei Stufen Rezeption von vorherigen VERA8-Ergebnissen [REPZ], deren Reflexion [RFL] und das Bestreben, daraus ggf. Unterrichtsveränderungen abzuleiten [VEA], gebildet wurden. Die Skalen zum Zyklusmodell traten somit an die Stelle der Attribution der Testleistung. Zusätzlich wurde mit der Unterstützung durch die Schulleitung eine weitere mögliche Ressource integriert, die als wichtige Voraussetzung gilt, das bereitgestellte Feedback zu nutzen.

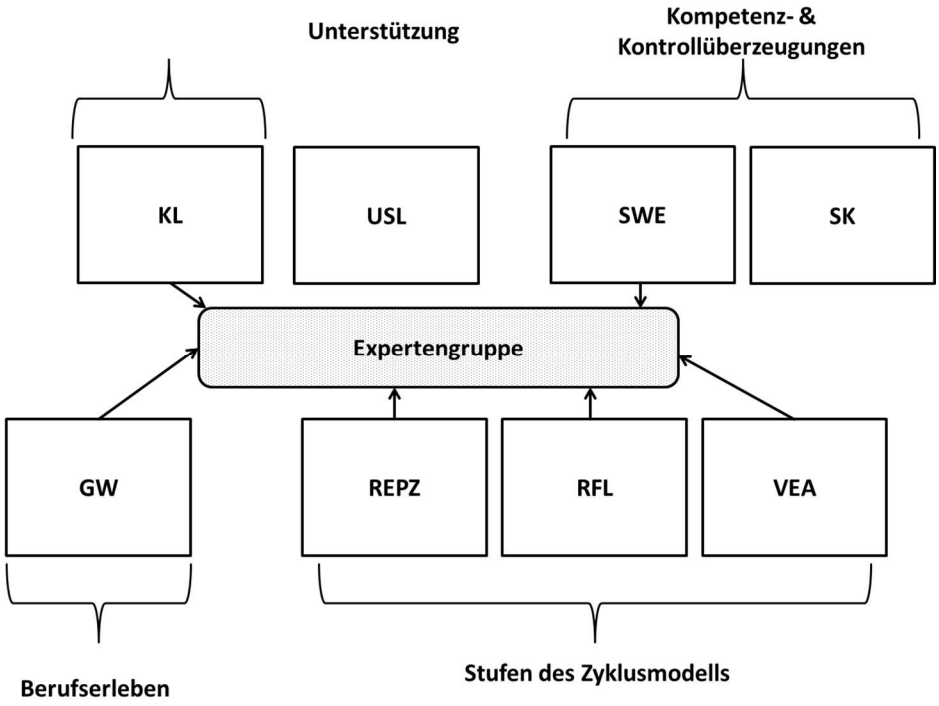


Abbildung 6.5: Modell M22 – allgemein feedback-relevante Überzeugungen als Prädiktoren des Vorbereitungsverhaltens

Auf Grundlage der fünf hier skizzierten Modelle wurde das Vorbereitungsverhalten nach den jeweiligen Klassen differenziert analysiert. Bevor die Ergebnisse dieses zweiten Analyseschritts dargestellt werden können, gibt das nachfolgende Kapitel zuerst aber eine globale Übersicht über die Ergebnisse der schriftlichen Befragung in beiden Studien.

7 Ergebnisse

Die in diesem Kapitel dargestellten Ergebnisse sind von verschiedener Natur. In Abschnitt 7.1 werden globale Ergebnisse der beiden Studien A und B zum Testcoaching im Zusammenhang mit VERA8 Mathematik des Jahres 2010 dargestellt. Die Darstellung bleibt weitestgehend auf einer Ebene, die ausschließlich deskriptiv wahrzunehmen ist. Vereinzelt werden allerdings in diesem Kapitel noch spezielle Hypothesen formuliert und die zugehörigen Befunde werden anschließend direkt unter Rückgriffe auf die unmittelbaren Ergebnisse diskutiert. Dies trifft u.a. auf die Unterscheidung zwischen Lehrkräften mit und ohne VERA8-Erfahrung zu (7.1.5). In Abschnitt 7.2 werden dann für die verschiedenen Modelle die Ergebnisse der latenten Klassenanalysen dargestellt und die wichtigsten Klassenlösungen interpretiert. Dies dient als Grundlage für weitergehende Analysen, in denen das Vorbereitungsverhalten mit den besten Modellen verbunden wird (7.3) und (7.4). Für die Diskussion der Ergebnisse vor der Folie der Forschungsfragen und eigentlichen Hypothesen dieser Arbeit wird auf Kapitel 8 verwiesen.

7.1 Qualität und Umfang der Vorbereitung auf VERA8

Im ersten Abschnitt werden die Ergebnisse aus den Studien A und B zu Umfang und Qualität der Vorbereitung beschrieben. Dabei werden für beide Studien die Ergebnisse von Fragebögen mit Haftklebezettel (Studien A1 und B1) und ohne Haftklebezettel (Studien A2 und B2) getrennt dargestellt. Diese Unterscheidung ist an dieser Stelle nötig, da auch bei diesen beiden Studien die Haftklebezettel einen positiven Effekt auf die Rücklaufquote aufwies und – wie nachfolgend deutlich wird – das Abbild des Vorbereitungsverhaltens beeinflussten. Der Abschnitt beginnt mit den Ergebnissen zum zeitlichen Umfang der Vorbereitung (7.1.1). Anschließend folgen die Ergebnisse zu den drei Testcoachingzugängen Familiarity Approach (FA), Content Approach (CA) und Test Wiseness Approach (TWS). Danach wird dargestellt, inwieweit sich Lehrkräfte mit und ohne Erfahrung mit VERA8 im zeitlichen Umfang und der Vorbereitungsgestaltung unterschieden (7.1.5). Es folgt die Betrachtung einer möglichen Schwerpunktsetzung durch die Lehrkräfte im Unterricht (7.1.6) und anschließend wird die Nutzung von Vorbereitungs- und Kompetenzheften betrachtet (7.1.7). Den Schluss bilden die Ergebnisse zur außerunterrichtlichen Vorbereitung auf VERA8 (7.1.8).

7.1.1 Zeitlicher Umfang der Vorbereitung

Der Umfang des Vorbereitungsverhaltens drückt sich innerhalb der Befragung durch die für die Vorbereitung aufgewendeten Unterrichtsstunden sowie die Anzahl der umgesetzten Maßnahmen aus. Über alle vier Teilstudien hinweg betrug die durchschnittliche Dauer der konkreten Vorbereitung auf die zentralen Vergleichsarbeiten zwei bis zweieinhalb Schulwochen, n=675, M=7.47, SE=0.21, SD=5.32, 95% CI=[7.04, 7.87] [120]. Bei einem vorgesehenen Stundenvolumen für Mathematik in der achten Jahrgangsstufe an Gymnasien in NRW von drei bis vier Wochenstunden bedeutet dies, dass Lehrkräfte die Unterrichtszeit von zwei bis drei Wochen für die Vorbereitung nutzten.

Tabelle 7.1

Anzahl der für die Vorbereitung aufgewendeten Unterrichtsstunden

Studie	Teilstudie	n	M (SD)	95% CI
A	A1	189	6.47 (4.99)	[5.79, 7.21]
	A2	182	8.45 (6.16)	[7.59, 9.35]
B	B1	206	7.20 (4.73)	[6.54, 7.84]
	B2	97	8.16 (5.12)	[7.13, 9.21]
Studien A1 & B1: mit Haftklebezettel			Studien A2 & B2: ohne Haftklebezettel	

Wie Tab. 7.1 zu entnehmen ist, zeigt sich der hohe Umfang, den Lehrkräfte als Stundenrahmen für die Vorbereitung ihrer Schüler auf die zentralen Vergleichsarbeiten angegeben haben, in allen vier Teilstudien und kann zumindest für ein Drittel der Grundgesamtheit als sicher gelten.[121] Zwischen den Teilstudien A1 und A2 besteht aber ein signifikanter Unterschied, t(348.09)=3.41, df=, p=.01[122], r=.17. Für die Teilstudien B1 und B2

[120] Die Konfidenzintervalle beruhen auf 1000 Bootstrapstichproben.

[121] Im Vergleich zwischen den Teilstudien mit Klebehaftzettel (A1 und B1) und den Teilstudien ohne Klebehaftzettel (A2 und B2) berichteten Lehrkräfte jeweils ein größeres Stundenvolumen in den beiden Teilstudien mit geringerem prozentualem Rücklauf (A2 und B2). Möglicherweise haben gerade die Lehrkräfte an der Studie teilgenommen, die eine mehrere Stunden umfassende Vorbereitung durchgeführt haben. Dem steht aber wiederum gegenüber, dass der Unterschied nur für die Studie A signifikant (p=.01) ist und eine Effektstärke im unteren Bereich aufweist.

[122] Die Verteilung der aufgewendeten Unterrichtsstunden weist keine Varianzhomogenität zwischen den beiden Teilstudien A1 und A2 auf (Levene-Test zu p=0.05) und folgt zusätzlich keiner Normalverteilung (getestet mit dem Kolmogorov-Smirnov-Test zu Niveau p=0.01).

ist der Unterschied hingegen nicht signifikant, t(300)=1.889, p=.03 [123] . Der hohe Stundenumfang wird auch in den nachfolgenden Histogrammen (Abb. 7.1 bis Abb. 7.4) noch einmal deutlich, in denen jeweils die Angaben zu den aufgewendeten Unterrichtsstunden in Schulwochen à 4 Wochenstunden zusammengefasst ist. Außerdem ist abgetragen, wie viele Lehrkräfte maximal zwei Unterrichtsstunden vorbereiteten. Eine Vorbereitung von maximal zwei Stunden entspricht dabei dem Umfang, der für einen möglicherweise beabsichtigten Familiarity Approach sinnvoll wäre. Alle darüber hinaus aufgewendete Unterrichtszeit ist nicht im Sinne des Instruments zentraler Vergleichsarbeiten und gefährdet ihre Funktionalität.

Bei der Betrachtung der Histogramm über alle vier Studien gemeinsam fällt auf, dass maximal ein Sechstel der Lehrkräfte nur in dem Rahmen vorbereitete, der ausreicht, um sich mit den Testinstrumenten vertraut zu machen, und von dem angenommen wird, dass er zur Steigerung der Testvalidität benötigt wird. In keiner Teilstudie gaben mehr als vier Befragte an, gar nicht vorzubereiten. Der Großteil der Lehrkräfte bereitete mindestens eine vollständige Woche vor, in drei von vier Teilstudien berichteten die Lehrkräfte mehrheitlich sogar von einem Umfang, der bis zu zwei volle Wochen beträgt. In allen vier Teilstudien bereiteten zumindest ein Viertel der Lehrkräfte sogar mindestens neun Stunden und in wiederum drei Teilstudien sogar zehn Prozent mehr als zwölf Stunden vor. Da die maximalen Stundenzahlen, die in den vier Teilstudien angegeben wurden bei (40|40|32|24) (Reihenfolge wie oben) liegen, kann davon ausgegangen werden, dass die befragten Lehrkräfte die zur Vorbereitung aufgewendeten Stunden qualitativ von einer Implementation von durch die zentralen Vergleichsarbeiten illustrierten Aufgaben oder Inhalte der Bildungsstandards trennen und es sich tatsächlich um eine gezielte Vorbereitung handelt.

[123] Die Verteilung der aufgewendeten Unterrichtsstunden besitzt für die beiden Teilstudien gleiche Varianz (H_0: Varianz ist gleich kann auf p=0.05 nicht abgelehnt werden). Die Verteilung folgt keiner Normalverteilung.

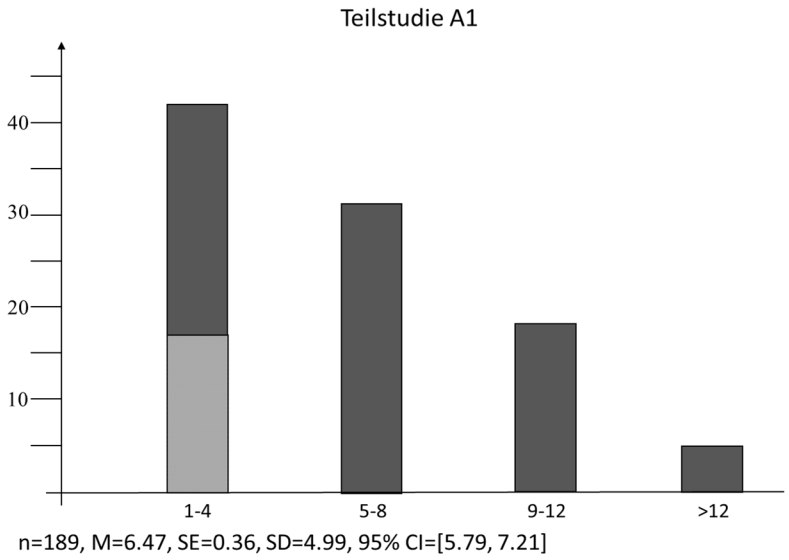


Abbildung 7.1: Aufgewendetes Stundenvolumen nach Schulwochen zusammengefasst – Teilstudie A1 (Verteilung der Lehrkräfte in Prozent – heller Anteil Lehrkräfte mit max. zweistündiger Vorbereitung)

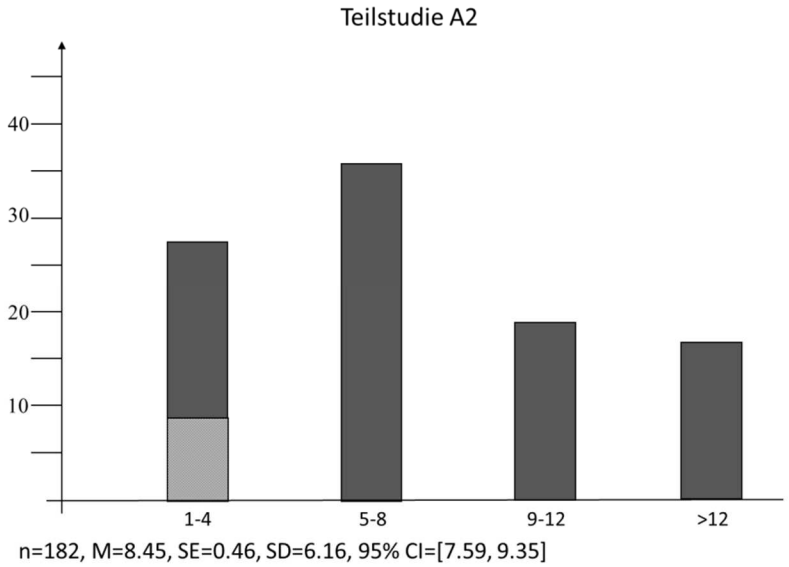


Abbildung 7.2: Aufgewendetes Stundenvolumen nach Schulwochen zusammengefasst– Teilstudie A2 (Verteilung der Lehrkräfte in Prozent – heller Anteil Lehrkräfte mit max. zweistündiger Vorbereitung)

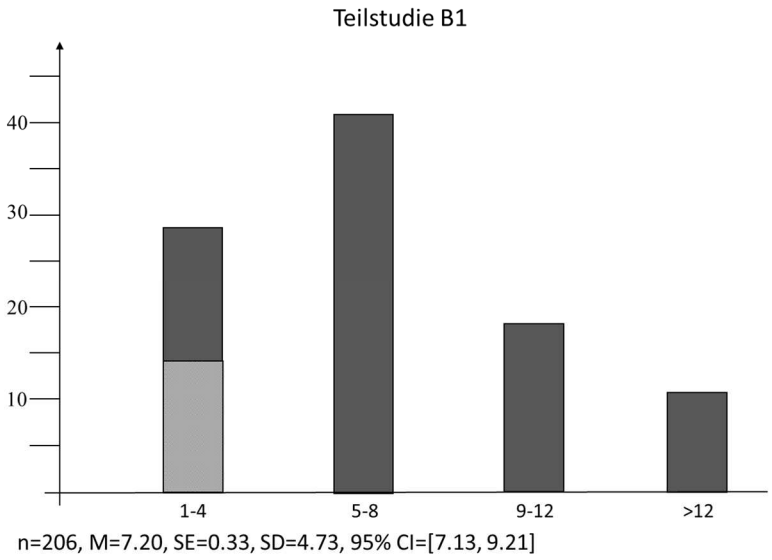


Abbildung 7.3: Aufgewendetes Stundenvolumen nach Schulwochen zusammengefasst– Teilstudie B1 (Verteilung der Lehrkräfte in Prozent – heller Anteil Lehrkräfte mit max. zweistündiger Vorbereitung)

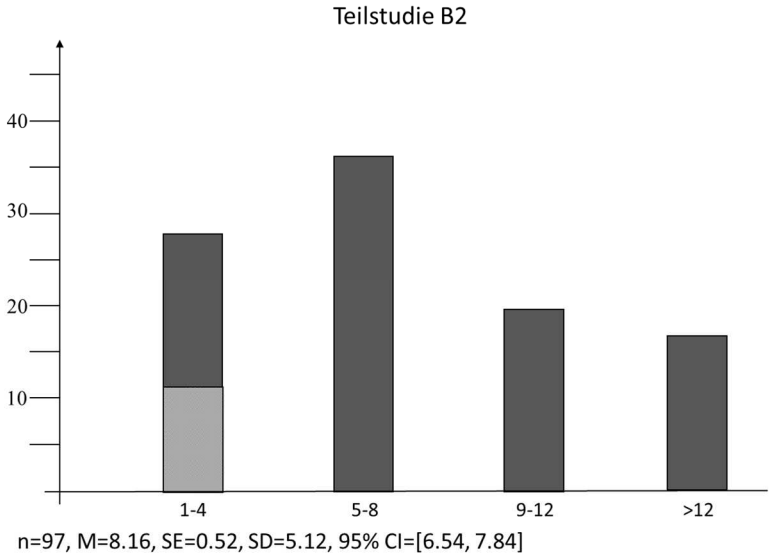


Abbildung 7.4: Aufgewendetes Stundenvolumen nach Schulwochen zusammengefasst– Teilstudie B2 (Verteilung der Lehrkräfte in Prozent – heller Anteil Lehrkräfte mit max. zweistündiger Vorbereitung)

7.1.2 Familiarity Approach

Der zweite Indikator für den Umfang der Vorbereitung ist die Anzahl der durchgeführten Maßnahmen. Insgesamt standen 13 vorgegebene Maßnahmen (als „Optionen“ bezeichnet) im Fragebogen zur Auswahl. Die Lehrkräfte wurden gebeten, diejenigen Maßnahmen anzugeben, die sie durchgeführt haben. Bei den ersten acht Maßnahmen handelt es sich um Möglichkeiten, die Vertrautheit mit den Tests aus den zentralen Lernstandserhebungen zu erhöhen und ein Familiarity Approach durchzuführen. Fünf Maßnahmen

(a) mit Testaufgaben früherer LSE üben lassen,

(b) alte Testaufgaben zur Verfügung gestellt,

(c) zu LSE ähnliche Aufgaben in vorherige Klassenarbeiten eingebaut,

(d) alte LSE-Aufgaben in Klassenarbeit eingebaut und

(e) Beispielaufgaben von der offizielle Homepage lösen lassen

beruhen dabei auf der Bearbeitung von Aufgaben, die in dem Format der Tests aus den zentralen Vergleichsarbeiten gestellt sind. Durch die Bearbeitung dieser Aufgaben werden somit nicht nur die prognostizierten Inhalte der zentralen Vergleichsarbeiten wiederholt und vertieft, sondern Lehrkräfte tragen durch den Einsatz der besonderen Art Rechnung, in der die Aufgaben gestellt sind (z.B. im Multiple-Choice-Format).

Bei den fünf Maßnahmen mit Aufgabenformatbezug zeigen sich deutliche Unterschiede zwischen der Nutzungshäufigkeit. Die größte Nutzungshäufigkeit besitzt (a) das Übenlassen mit alten LSE-Aufgaben. In allen vier Teilstudien haben dies mindestens achtzig Prozent der Lehrkräfte angegeben (A1: 80.1%, A2: 85.7%, B1: 86.9%, B2: 86.4%). Schon deutlich weniger, aber immer noch von weit mehr als über der Hälfte der Lehrkräfte wurde angegeben, dass sie den Schülern Aufgaben zur eigenständigen Bearbeitung zur Verfügung gestellt haben (b). Dabei gibt es signifikante Unterschiede zwischen den Teilstudien der Studie A (A1 nur 57.6%, A2: 71.4%, $\chi^2[3]=10.981$, $p=.01$, $w=.17$), die sich mit Blick auf die Angaben in Studie B (B1: 69.5% und B2: 60.2%) nicht durch die Rücklaufquote erklären lassen. Knapp dreißig Prozent der Lehrkräfte haben zu den LSE-Aufgaben ähnliche Aufgaben in vorherige Klassenarbeiten einfließen lassen, nur zwischen 7.9% (A1 und B2) und 13.8% der Befragten haben dazu aber originale Aufgaben verwendet. Ebenfalls dreißig Prozent der Lehrkräfte haben außerdem im Unterricht Aufgaben lösen lassen, die auf den offiziellen Homepages des Instituts für Qualität im Bildungswesen (IQB) oder des Ministeriums für Schule und Weiterbildung (MSW) zur Illustration der inhaltlichen Schwerpunkte einsehbar waren. Insgesamt haben im

Durchschnitt neun von zehn Lehrkräften mit LSE-Aufgaben oder mit zu ihnen ähnlichen Aufgaben in ihrer Klasse auf VERA8 vorbereitet.

Tabelle 7.2

Welche der folgenden Optionen haben Sie in dieser Klasse als unmittelbare Vorbereitung auf die LSE genutzt? – Familiarity Approach-Maßnahmen differenziert nach Teilstudien

Maßnahme	Studie A1 (n=191) absolut (%)	Studie A2 (n=189) absolut (%)	Studie B1 (n=213) absolut (%)	Studie B2 (n=103) absolut (%)
(a) mit Testaufgaben früherer LSE üben lassen	153 (80.1%)	162 (85.7%)	185 (86.9%)	89 (86.4%)
(b) alte Testaufgaben zur Verfügung gestellt	110 (57.6%)	135 (71.4%)	148 (69.5%)	62 (60.2%)
(c) zu LSE ähnliche Aufgaben in vorherige Klassenarbeiten eingebaut	57 (29.8%)	64 (33.9%)	60 (28.2%)	30 (29.1%)
(d) alte LSE-Aufgaben in Klassenarbeiten	15 (7.9%)	26 (13.8%)	25 (11.7%)	15 (7.9%)
(e) Beispielaufgaben von Homepage lösen lassen	44 (23.0%)	51 (27.0%)	54 (25.4%)	29 (28.2%)
Familiarity Approach mit Aufgaben insgesamt	172 (90.1%)	177 (93.7%)	199 (93.4%)	95 (92.2%)
(f) in der Klasse gemeinsam die offizielle Internetseite besucht	7 (3.7%)	17 (9.0%)	16 (7.5%)	5 (4.9%)
(g) Schüler auf offizielle Internetseite hingewiesen	114 (59.7%)	99 (52.4%)	108 (50.7%)	42 (40.8%)
(h) Testsituation simuliert	36 (18.8%)	43 (22.8%)	39 (18.3%)	15 (14.6%)

Wichtig im Sinne eines Familiarity Approach ist auch, sich über die Abläufe und Ziele des Tests zu informieren. Dazu bieten sich Informationsmaterialien an, die auf der länderübergreifenden Internetpräsenz des IQB und spezifisch auf den Internetseiten der Bundesländer wie für Nordrhein-Westfalen die des MSW verfügbar sind. Auch ist es

denkbar, die Testsituation komplett innerhalb der Klasse zu simulieren. Während das Üben mit alten Testaufgaben häufiger im Klassenraum stattfand als es als Empfehlung für das häusliche Üben genannt wurde, stellte sich dies beim Besuch der offiziellen Internetseiten umgekehrt da: Nur maximal neun Prozent der Lehrkräfte hat zumindest eine der offiziellen Internetseiten gemeinsam im Klassenverbund besucht (f), aber zwischen 59.7% (A1) und 40.8% (B2) der Lehrkräfte haben den Besuch ihren Schülern anempfohlen (g). Als Erklärung kann hier gelten, dass Arbeiten am Computer bzw. Erkundigungen im Internet mit erhöhtem Aufwand (beispielsweise einem Raumwechsel) verbunden sind und andersherum die meisten der Gymnasiasten zu Hause über einen Internetzugang verfügen werden. Von einer Simulation der Testsituation haben nach eigenen Angaben zwischen 22.8% (A2) und 14.6% (B2) der Lehrkräfte Gebrauch gemacht.

Bis auf Maßnahme (g) handelt es sich immer um Maßnahmen, die im Unterricht durchgeführt werden müssen (a, b, c, d, e, h) oder zumindest im Unterricht durchgeführt werden können (f). Die Zahl der verschiedenen durchgeführten Maßnahmen kann dadurch auch eine Art Vorbereitungsintensität ausdrücken. Wenngleich eine höhere Maßnahmenvariabilität nicht zwangsweise zu mehr aufgewendeter Unterrichtszeit führt, beansprucht sie doch zumindest mehr Vorbereitungszeit und drückt gleichzeitig ein größeres Interesse an einer umfassenden Vorbereitung aus. Abschließend wurde dementsprechend für diese sieben Maßnahmen im Sinne eines Familiarity Approach ein Summenindex gebildet, der in nachfolgenden vier Histogrammen die Nutzungsvariabilität angibt (Abb 7.5 bis 7.8).

Den Histogrammen kann entnommen werden, dass die Nutzungsvariabilität gemessen durch die sieben FA-Maßnahmen nur sehr gering ausfällt. In allen vier Teilstudien geben ca. fünfzig Prozent der Befragten an, nur zwei verschiedene Maßnahmen genutzt zu haben. Der Modalwert liegt dreimal bei zwei Maßnahmen (A1, A2 und B2) und nur einmal bei drei Maßnahmen (B1). Das Maximum an genutzten Maßnahmen liegt zwischen fünf (A1) und sieben (B1). Insgesamt hat nur eine Lehrkraft berichtet, tatsächlich alle sieben Maßnahmen durchgeführt zu haben.

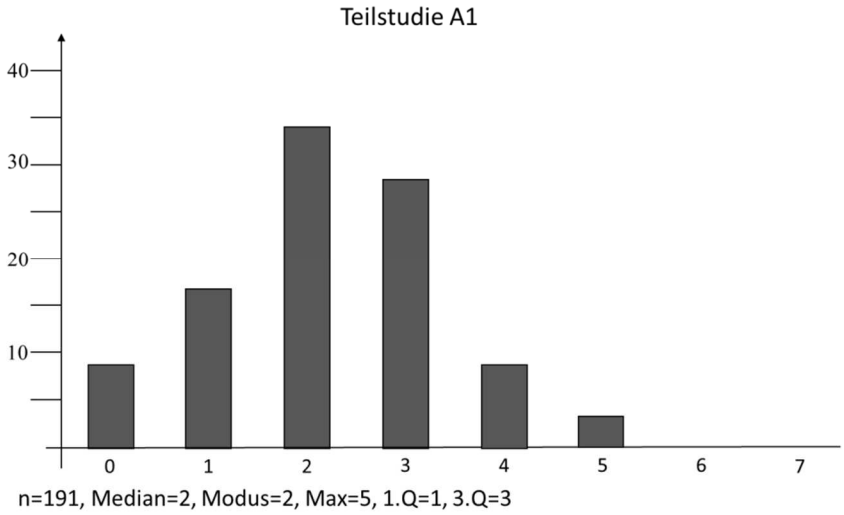


Abbildung 7.5: Darstellung der Nutzungsvariabilität von FA-Maßnahmen in Teilstudie A1 (Verteilung der Lehrkräfte in Prozent)

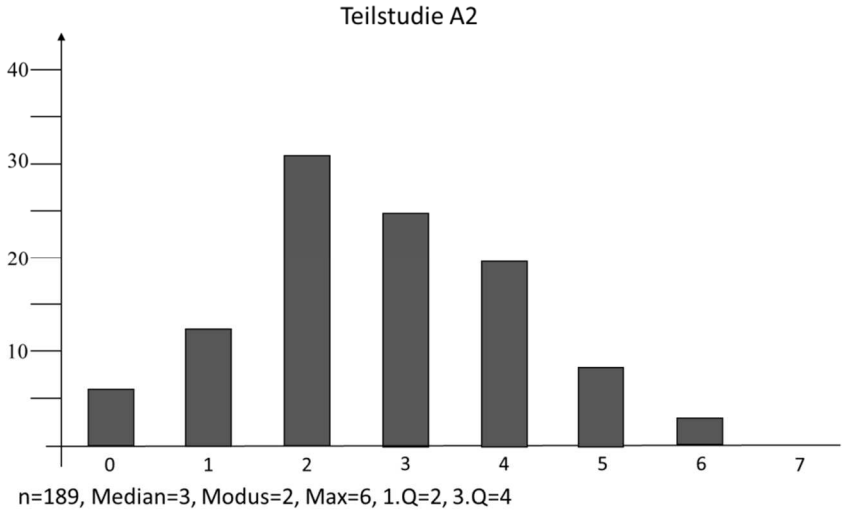


Abbildung 7.6: Darstellung der Nutzungsvariabilität von FA-Maßnahmen in Teilstudie A2 (Verteilung der Lehrkräfte in Prozent)

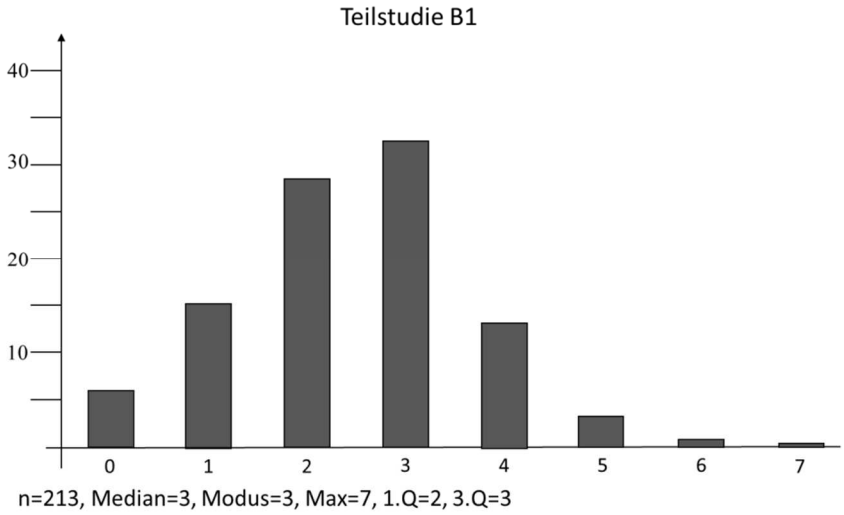


Abbildung 7.7: Darstellung der Nutzungsvariabilität von FA-Maßnahmen in Teilstudie B1 (Verteilung der Lehrkräfte in Prozent)

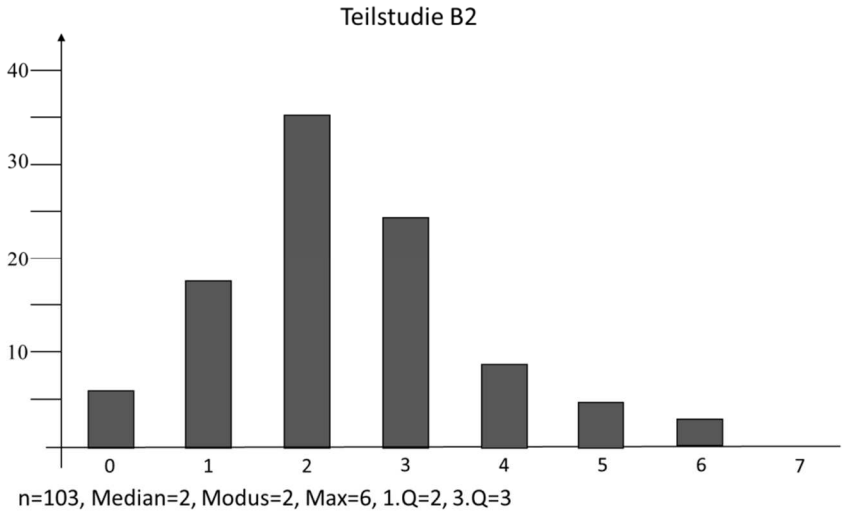


Abbildung 7.8: Darstellung der Nutzungsvariabilität von FA-Maßnahmen in Teilstudie B2 (Verteilung der Lehrkräfte in Prozent)

Neben den genannten Maßnahmen kann man auch im Sinne eines Familiarity Approachs bestimmte Bestandteile der Aufgaben thematisieren. Dies ist nicht an konkrete Maßnahmen gebunden. Im Fragebogen wurde daher auch erhoben, ob Lehrkräfte die Aufgabenstellungen ((p) „wie man die Aufgabenstellung der LSE-Aufgaben richtig versteht“ und (q) „wie man den Aufgabenstellungen der LSE-Aufgaben die wichtigen Informationen entnimmt“) und Antwortmöglichkeiten ((r) „wie man in Hinblick auf die speziellen Antwortformate richtig

antwortet“) angesprochen haben. Diese Themen anzusprechen, unterstützt den Aufbau einer Testkompetenz, die unabhängig von zentralen Vergleichsarbeiten bei der Bearbeitung von Leistungstests hilfreich ist und deren Besitz eine notwendige Voraussetzung für die maximal Testvalidität darstellt. Da dieser Teil einer Testkompetenz nicht an konkrete Maßnahmen gekoppelt ist, bedarf es einiger Reflexionsleistung durch die Lehrkraft, um die Themen mit den Schülern zu besprechen; (q) geht hier sogar noch darüber hinaus, da die Informationsentnahme auch als Teil des im Zusammenhang mit internationaler Schulleistungsmessung wie PISA und PIRLS diskutierten Literacy-Konzepts gesehen werden kann. Testaufgaben sind in diesem Sinne eine sehr spezielle Form eines Sachtexts. Dieser Teil der Testkompetenz ist folglich noch einmal anspruchsvoller als die beiden anderen Themen und stellt wahrscheinlich auch für Lehrkräfte eine größere Herausforderung dar.

Im Fragebogen wurde an dieser Stelle unterschieden zwischen der mehrfachen Thematisierung (M), der einmaligen Thematisierung (E) und keiner Thematisierung (K). Auf Aggregatebene zeigt sich für alle drei Themen und in allen vier Teilstudien, dass ungefähr die Hälfte der Lehrkräfte die drei Themen mehrfach angesprochen haben. Zwischen weiteren 26.2% (A1 und B2) und 18.5% (A2) haben das richtige Verständnis der Aufgabenstellung (p) zumindest einmal thematisiert und zwischen 26.5% (A2) und 22.1% (B1) haben zumindest einmal angesprochen, wie man bei den speziellen Antwortformaten richtig antwortet (r). Die Informationsentnahme aus der Aufgabenstellung (q) wurde in allen vier Teilstudien von minimal mehr als zwanzig Prozent der Lehrkräfte gar nicht angesprochen. Dies ist der höchste Wert, bei den beiden anderen Themen (die ausschließlich Bezug zu den VERA8-Tests aufweisen) lag der Wert immer unter zwanzig Prozent. Dabei setzt sich der Wert für keinerlei Thematisierung von (q) sowohl aus einer im Vergleich zu (p) und (r) geringerer Anzahl von Lehrkräften zusammen, die (q) mehrfach angesprochen haben, als auch aus einer geringer Anzahl von Lehrkräften, die dies zumindest einmalig thematisierten.

Tabelle 7.3

Welche der folgenden Themen haben Sie im Unterricht angesprochen? – differenziert nach Teilstudien

	Studie A1 (n=191)			Studie A2 (n=189)			Studie B1 (n=213)			Studie B2 (n=103)		
Maßnahme	mehrfach (%)	einmal (%)	gar nicht (%)	mehrfach (%)	einmal (%)	gar nicht (%)	mehrfach (%)	einmal (%)	gar nicht (%)	mehrfach (%)	einmal (%)	gar nicht (%)
(p) wie man die Aufgabenstellung der LSE-Aufgaben richtig versteht	93 (48.7%)	50 (26.2%)	37 (19.4%)	106 (56.1%)	35 (18.5%)	42 (22.2%)	120 (56.3%)	46 (21.6%)	39 (18.3%)	56 (54.4%)	27 (26.2%)	19 (18.4%)
(q) wie man den Aufgabenstellungen der LSE-Aufgaben die wichtigen Informationen entnimmt	93 (48.7%)	38 (19.9%)	48 (24.1%)	96 (50.8%)	34 (18.0%)	50 (26.5%)	114 (53.5%)	40 (18.8%)	50 (23.4%)	53 (51.5%)	24 (23.3%)	21 (20.4%)
(r) wie man in Hinblick auf die speziellen Antwortformate richtig antwortet	94 (49.2%)	49 (25.7%)	36 (18.8%)	100 (52.9%)	50 (26.5%)	36 (19.0%)	121 (56.8%)	47 (22.1%)	37 (17.4%)	58 (56.3%)	23 (22.3%)	19 (18.4%)

Anmerkung: Die oben zu den Teilstudien gegebene Anzahl bezieht sich immer auf die Anzahl der Lehrkräfte, zu denen mindestens zu einer der Optionen eine Angabe vorlag. Es können sich daher für die einzelnen Optionen Abweichungen in der Summe ergeben.

7.1.3 Content Approach

Weitere fünf der abgefragten Maßnahmen lassen sich einem Content Approach (CA) zurechnen. Im Rahmen der vier Teilstudien wurde in Bezug auf die Kernlehrpläne Mathematik NRW zwischen Inhaltskompetenzen und Prozesskompetenzen unterschieden und berücksichtigt, dass in den Jahren 2004 bis 2008 jeweils nur eine Prozesskompetenz schwerpunktmäßig erhoben wurde, während immer alle vier Inhaltskompetenzen erhoben wurden. Bei den Maßnahmen handelt es sich um:

(i) *alle Inhaltsbereiche* noch einmal im Klassenverbund zu wiederholen,

(j) den *Schülern* zu *empfehlen*, zu Hause noch einmal alle Inhaltsbereiche zu wiederholen,

(k) alle Prozessbereiche noch einmal zu wiederholen,

(l) den *Schülern* zu *empfehlen*, zu Hause noch einmal alle Inhaltsbereiche zu wiederholen und

(m) *eine Prozesskompetenz* besonders üben zu lassen.

Die Maßnahmen (i) und (j) entsprechen dem klassischen Vorgehen einer Vorbereitung auf Klassenarbeiten und (Abitur-)Klausuren in Mathematik. Alle für den Test relevanten Inhaltsbereiche werden noch einmal wiederholt. In allen vier Teilstudien wurden diese beiden Maßnahmen folglich auch häufiger berichtet als Übungs- und Wiederholungsphasen zu Prozesskompetenzen. Von den Lehrkräften berichteten zwischen 37.2% (A1) und 47.1% (A2), dass sie mit ihren Schülern alle Inhaltsbereiche wiederholt haben, während die Mehrheit die durchschnittlich relativ umfangreiche Vorbereitungszeit offensichtlich nicht auf die Inhaltsbereiche ausgerichtet genutzt hat. Ähnlich verhält es sich mit dem Hinweis an die Schüler, die Inhaltsbereiche alle selbstständig zu wiederholen. Hier von berichteten etwas mehr Lehrkräfte, konkret zwischen 45.5% (A1) und 57.3% (B1). Zumindest eine der beiden Varianten umgesetzt haben zwischen 64.9% (A1) und 71.8% (B2) der Befragten. Trotzdem zeigen sich unerwartete niedrige Werte für die Inhaltsbereiche, der inklusive der empfohlenen außerschulischen Vorbereitung unter dem Wert für das Üben mit früheren LSE-Aufgaben liegt.

Die Prozesskompetenzen sind erst im Zuge der neu formulierten Kernlehrpläne sichtbarer Bestandteil des intendierten Curriculums geworden. Möglicherweise sind diese daher noch nicht in gleicher Weise ins Bewusstsein der Lehrkräfte gelangt. Andererseits wurde ihre Bedeutung durch die Schwerpunktsetzung in den ersten vier Durchgängen der zentralen Vergleichsarbeiten als Lernstand9 bzw. Lernstand8 in NRW hervorgehoben. Beide Überlegungen ließen daher Abweichungen auch in der Implementation erwarten (erstes

nach unten, zweites nach oben), die sich auch in die Vorbereitung auswirken. Tatsächlich zeigt sich in den vier Teilstudien für die Maßnahmen (k) und (l), alle Prozessbereiche zu wiederholen bzw. dies den Schülern empfohlen zu haben, deutlich geringere Umsetzungen durch die befragten Lehrkräfte. Nur zwischen 24.1% (A1) und 28.2% (B2) haben selbst im Unterricht alle Prozesskompetenzen wiederholt und nur zwischen 23.8% (A2) und 32.0% (B2) haben dies ihren Schülern nahe gelegt. Noch geringe fallen die Angaben für das gezielte Üben einer Prozesskompetenz aus, das nur zwischen 8.5% (B1) und 11.6% (A2) der Lehrkräfte berichteten. Dies muss vor dem Hintergrund betrachtet werden, dass seit dem Jahr 2009 keine Schwerpunktsetzung für VERA8 in Mathematik vorgenommen wird und somit eine Schwerpunktsetzung in der Vorbereitung nur insofern sinnvoll erscheint, wenn vorab ein Defizit in diesem Bereich deutlich wurde. Betrachtet man wiederum die drei Maßnahmen gemeinsam, so ergaben sich Werte zwischen 45.5% (A1) und 47.6% (B2). Fasst man nun alle fünf Maßnahmen für ein Content Approach zusammen, ergeben sich nur minimal größere Werte als allein für die auf die Inhaltsbereiche bezogenen Maßnahmen (A1: 72.3%, A2: 75.1%, B1: 75.1%, B2: 76.7%). Es bereiteten folglich also gerade diejenigen Lehrkräfte zielgerichtet auf die Prozesskompetenzen vor, die auch alle Inhaltsbereiche noch einmal wiederholten. Dabei sind die Unterschiede zwischen den vier Teilstudien vernachlässigbar gering.

Tabelle 7.4

Welche der folgenden Optionen haben Sie in dieser Klasse als unmittelbare Vorbereitung auf die LSE genutzt? – CA-Maßnahmen differenziert nach Teilstudien

Maßnahme	Studie A1 (n=191) absolut (%)	Studie A2 (n=189) absolut (%)	Studie B1 (n=213) absolut (%)	Studie B2 (n=103) Absolut (%)
(i) *alle* Inhaltsbereiche wiederholt	71 (37.2%)	89 (47.1%)	80 (37.6%)	47 (45.6%)
(j) *empfohlen,* alle Inhaltsbereiche zu wiederholen	87 (45.5%)	95 (50.3%)	122 (57.3%)	58 (56.3%)
(i)+(j)	124 (64.9%)	134 (70.9%)	148 (69.5%)	74 (71.8%)
(k) *alle* Prozessbereiche wiederholt	46 (24.1%)	52 (27.5%)	55 (25.8%)	29 (28.2%)
(l) *empfohlen,* alle Prozesskompetenzen zu wiederholen	46 (24.1%)	45 (23.8%)	66 (31.0%)	33 (32.0%)
(m) *eine* Prozesskompetenz besonders üben lassen	26 (13.6%)	22 (11.6%)	18 (8.5%)	12 (11.7%)

Maßnahme	Studie A1 (n=191) absolut (%)	Studie A2 (n=189) absolut (%)	Studie B1 (n=213) absolut (%)	Studie B2 (n=103) Absolut (%)
(k)+(l)+(m)	87 (45.5%)	88 (46.6%)	100 (46.9%)	49 (47.6%)
(vorgegebene) Content Approach Maßnahmen insgesamt	138 (72.3%)	142 (75.1%)	160 (75.1%)	79 (76.7%)

Außer den 13 vorformulierten Maßnahmen konnten auch Maßnahmen frei formuliert werden. 56 Lehrkräfte gaben an, zusätzlich andere Maßnahmen zur Vorbereitung genutzt zu haben. Sie setzten selbst entwickelte Tests oder Quizspiele zur Diagnose oder zur Übung ein, ließen Vorbereitungsaufgaben aus dem grundständigen Schulbuch oder von anderen als den offiziellen Internetseiten bearbeiten, führten mit der Wochenplanmethode eine Übungsphase durch oder wiederholten ganz gezielt einzelne Teilbereiche wie den Umgang mit Excel oder die Prozentrechnung. Auch berichteten Lehrkräfte vereinzelt über die Teilnahme an Wettbewerben wie dem Känguru-Wettbewerb und an einigen Schulen wurde ein Projekttag für die Vorbereitung genutzt. Referate über Teilbereiche oder Fragestunde wurden ebenfalls als weitere Maßnahmen genannt. Die frei formulierten Maßnahmen wurden in Maßnahmen im Sinne eines Familiarity Approach und im Sinne eines Content Approach kodiert. Verbindet man diese Maßnahmen mit den vorgegebenen Maßnahmen, erhöhen sich die Werte für die Nutzung von Familiarity Approach-Maßnahmen und Content Approach-Maßnahmen leicht, wobei der Anstieg für die einem Content Approach zugeordneten Maßnahmen höher ist (vgl. Tab. 7.5). Dieser Befund weist anders als die Nutzungsvariabilität der FA-Maßnahmen darauf hin, dass die aufgewendete Unterrichtszeit eher im Sinne eines FA genutzt wird. Wenngleich die frei formulierten CA-Maßnahmen mehrmalige Durchführungen nahelegen (z.B. Wochenplanarbeit), gaben ca. zwanzig Prozent der Lehrkräfte an, keine inhaltliche Vorbereitung durchgeführt zu haben.

Tabelle 7.5

Nutzung von vorgegebenen und frei formulierten Maßnahmen in der Vorbereitung – differenziert nach Teilstudien

Maßnahme	Studie A1 (n=191) absolut (%)	Studie A2 (n=189) absolut (%)	Studie B1 (n=213) absolut (%)	Studie B2 (n=103) Absolut (%)
Familiarity Approach mit Aufgaben: *vorformulierte Maßnahmen*	172 (90.1%)	177 (93.7%)	199 (93.4%)	95 (92.2%)
Familiarity Approach mit Aufgaben: *vorformulierte und frei formulierte Maßnahmen*	174 (91.1%)	177 (93.7%)	201 (94.4%)	95 (92.2%)
Content Approach: *vorformulierte Maßnahmen*	138 (72.3%)	142 (75.1%)	160 (75.1%)	79 (76.7%)
Content Approach: *vorformulierte und frei formulierte Maßnahmen*	145 (75.9%)	151 (79.9%)	166 (77.9%)	83 (80.6%)

7.1.4 Test Wiseness Approach

Die dritte Art Testcoaching zu betreiben, besteht in der Vermittlung von Tipps und Tricks, wie man die Lösungen der Testaufgaben besser erkennt oder generell mit Tests zurechtkommt. Diese Test-Wiseness-Strategien (TWS) können testspezifisch auf VERA8 zugeschnitten sein oder allgemein für Tests hilfreich sein. Durch die relativ geringe Anzahl an Testdurchläufen von VERA8 oder vorhergehende zentrale Lernstandserhebungen einerseits und durch die umfangreiche Kontrolle des Entwicklungsprozesses andererseits sind spezifische TWS als Gegenstand einer Vorbereitung auf VERA8 eher wenig sinnvoll. Wahrscheinlich ist es eher, dass Lehrkräfte auf allgemeine TWS zurückgreifen, wenn sie ihren Schülern solche Strategien vermitteln wollen. Es wurden daher neun allgemeine TWS vorformuliert und abgefragt, ob Lehrkräfte ihren Schülern diese mit auf den Weg gegeben haben.

Tabelle 7.6

Welche der folgenden Strategien haben Sie im Unterricht angesprochen? – differenziert nach Teilstudien

	Studie A1		Studie A2		Studie B1		Studie B2	
	Ja	Nein	Ja	Nein	Ja	Nein	Ja	Nein
Test-Wiseness-Strategie	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
(TWS1)…sich nicht zu lang an einer Frage aufzuhalten	128 (69.9%)	55 (30.1%)	144 (78.7%)	39 (21.2%)	155 (75.6%)	50 (24.4%)	68 (72.3%)	26 (27.7%)
(TWS2)…sich mit den Antworten vertraut zu machen.	115 (63.5)	66 (36.5%)	128 (69.9%)	55 (30.1%)	152 (74.15)	53 (25.9%)	72 (75.8%)	23 (24.2%)
(TWS3)…alle Antworten in Betracht zu ziehen, bevor man sich entscheidet.	112 (60.9%)	72 (39.1%)	118 (64.8%)	64 (35.2%)	149 (72.7%)	56 (27.3%)	62 (63.9%)	35 (36.1%)
(TWS4)…die Instruktionen und Fragen genau zu lesen.	172 (91.0%)	17 (9.0%)	176 (93.6%)	12 (6.4)	200 (94.8%)	11 (5.2%)	96 (96.0%)	4 (4.0%)
(TWS5)…bei Multiple-Choice-Fragen zu raten, wenn man die Antwort nicht weiß.	38 (21.0%)	143 (79.0%)	57 (31.7%)	123 (66.3%)	60 (29.9%)	141 (70.1%)	17 (18.3%)	76 (81.7%)
(TWS6)…zuerst die Fragen zu beantworten, bei denen man sich sicher fühlt.	127 (69.4%)	56 (30.6%)	137 (74.1%)	48 (25.9%)	154 (75.1%)	51 (24.9%)	66 (69.5%)	29 (30.5%)
(TWS7)…sich spontane Einfälle zu notieren.	32 (18.2%)	144 (81.8%)	20 (11.5%)	154 (88.5%)	30 (15.4%)	165 (84.5%)	10 (10.8%)	83 (89.2%)
(TWS8)…auf grammatikalische Einschränkungen der möglichen Antworten zu achten.	46 (25.6%)	134 (74.4%)	48 (25.4%)	129 (68.9%)	61 (30.5%)	139 (69.5%)	27 (29.0%)	66 (71.0%)
(TWS9)…im Zweifel die erste Idee zu wählen, weil dies meist das Beste ist.	7 (4.0%)	170 (96.0%)	6 (3.4%)	170 (96.6%)	10 (5.2%)	183 (94.8%)	3 (3.3%)	89 (96.7%)

Tab. 7.6 gibt für jede dieser neun Strategien an, wie hoch der Anteil an Lehrkräfte war, der seinen Schülern diese Strategien versuchte zu vermitteln. Bei den vorgegebenen TWS können drei Gruppen unterschieden werden. Die TWS2 („…sich mit den Antworten vertraut zu machen“), TWS3 („…alle Antworten in Betracht zu ziehen, bevor man sich entscheidet“) und TWS4 („…die Instruktionen und Fragen genau zu lesen“) zielen darauf ab, die Chance zu erhöhen, tatsächlich zu der richtigen Lösung im Sinne der Aufgabenintention zu gelangen. Die Strategien erhöhen ggf. den Testscore, sie ermöglichen außerdem gleichzeitig eine höhere Testvalidität, da die Konzentration der Testpersonen erhöht wird. TWS4 kann dabei

auf jede Testaufgabe angewendet werden, TWS2 und TWS3 sinn nur sinnvoll anzuwenden, wenn bereits mögliche Antworten zur Auswahl stehen, wie das vor allem bei Multiple-Choice-Aufgaben vorkommt. Entsprechend überrascht es auch nicht, wenn Lehrkräfte in allen vier Teilstudien jeweils in über neunzig Prozent der Fälle die TWS4 ihren Schülern mit auf den Weg gegeben haben (A1: 91,0%, A2: 93.6%, B1: 94.8%, B2: 96.0%), während die Werte für die beiden anderen Strategien etwas geringer ausfielen. Die TWS2 weist dabei noch leicht höhere Werte auf (zwischen 63.5% und 75.8%) als die TWS3 (zwischen 60.9% und 72.7%), welches sich dadurch erklären lässt, dass die beiden Strategien im Prinzip auf dasselbe (die Antwortmöglichkeiten) abzielen, aber TWS3 einen spezifischeren Umgang mit den dargebotenen Antwortmöglichkeiten verlangt, den man nicht als richtig ansehen muss.

Die zweite Gruppe von Strategien umfasst mit TWS5 („…bei Multiple-Choice-Fragen zu raten, wenn man die Antwort nicht weiß“) und TWS8 („…auf grammatikalische Einschränkungen der möglichen Antworten zu achten“) Strategien, die darauf abzielen, die richtige Lösung anzugeben. Es spielt dabei keine Rolle, ob man dafür tatsächlich auf die Kompetenz zurückgreift, die mit der Testaufgabe gemessen werden soll, oder anderes Wissen heranzieht. Im Gegensatz zur ersten Gruppe soll mit den Strategien folglich nur der Testscore gesteigert werden, die Sicherung der Testvalidität wird nicht angestrebt bzw. sogar untergraben. TWS5 wurde zwischen 18.3% (B2) und 31.7% (A2)[124] der Fälle angesprochen, TWS8 thematisierten die Lehrkräfte zu 25.4% (A2) bis 30.5% (B1). Lehrkräfte haben folglich diese beiden Strategien deutlich seltener in ihrem Unterricht angesprochen als die TWS aus der ersten Gruppe.[125]

Die Strategien TWS1 („…sich nicht zu lang an einer Frage aufzuhalten“), TWS6 („…zuerst die Fragen zu beantworten, bei denen man sich sicher fühlt“), TWS7 („…sich spontane Einfälle zu notieren“) und TWS9 („…im Zweifel die erste Idee zu wählen, weil die meist das Beste ist“)

bilden die dritte Gruppe der TWS. Sie zielen darauf ab, durch geschicktes Vorgehen die Testzeit optimal zu nutzen und in der begrenzten Zeit möglichst viele Lösungen zu finden. Ähnlich zur ersten Gruppe wird durch dieses rationale Bearbeiten eines Tests seine Validität erhöht, weil mehr Testaufgaben bearbeitet werden und die Bearbeitungsreihenfolge nicht durch die Darbietung der Reihenfolge im Test beeinflusst wird. In allen vier Teilstudien zeigt sich eine überraschende Diskrepanz zwischen den Strategien TWS1 und TWS6 auf der einen und der TWS7 sowie der TWS9 auf der anderen Seite. Während die TWS1 zwischen 69.9% (A1) und 78.7% (A2) Schülern mit auf den Weg gegeben wurde und sich die Werte für die WTS6 mit 69.4% in A1 als niedrigstem Wert und 75.1% in B1 als höchstem Wert in den vier Teilstudien ähnlich abbilden, liegen die Werte für die anderen beiden TWS deutlich darunter. Der höchste Wert liegt für die TWS7 bei 18.2% (A1), der niedrigste sogar nur bei 10.8% (B2)

[124] Bemerkenswert ist an dieser Stelle, dass es auf 5%-Niveau einen signifikanten Unterschied zwischen den Teilstudien gibt. Die TWS5 wurde in den Studien A1 und B2 signifikant weniger als thematisierte TWS genannt als in den Teilstudien A2 und B1.

[125] Trotzdem überrascht die hohe Zahl an Lehrkräften, die die TWS8 thematisierten, da grammatikalische Einschränkungen in Mathematik-Tests eigentlich keine Rolle spielen. Möglicherweise wurde das Item im Fragebogen missverstanden.

und die TWS9 gaben nur zwischen 5.2% (B1) und 3.3% (B2) der Lehrkräfte an, angesprochen zu haben. Eine mögliche Erklärung für diesen Unterschied zwischen den vier Strategien liegt darin, dass bei den TWS7 & 9 nicht direkt offensichtlich ist, wieso diese zu einer effizienteren Testbearbeitung führen.

31 Lehrkräfte haben außerdem TWS angegeben, die nicht unter den neun vorformulierten TWS waren. Elf der 32 genannten Strategien lassen sich auch in die drei Gruppen einordnen. Vier Strategien lassen sich der Gruppe 1 zu ordnen (Fragen mehrfach durchsehen, den außermathematischen Kontext genau beachten, Aufgabenstellung veranschaulichen, auch simpel erscheinende Aufgaben bearbeiten). Für die Gruppe 2 konnte eine Strategie (Ausschlussverfahren anwenden) fünfmal identifiziert werden. In die Gruppe 3 fallen fünf Strategien (sicher gelöste Aufgaben markieren und bei zweiter Durchsicht aussparen, nicht an Reihenfolge halten, schwierige Aufgaben zum Schluss bearbeiten, ausgelassene Aufgaben notieren, Ruhe bewahren). Weitere Strategien zeugen von einer anderen Wahrnehmung von VERA8, die bei vereinzelten Lehrkräften zu erkennen war (z.B. keine mathematische Genauigkeit in der Aufgabenstellung anzunehmen), oder zielen auf die Konzentrationsleistung der Schüler (z.B. Ruhe bewahren, sich Zeit für Überlegungen zu nehmen, Nebenrechnungen zu notieren).

7.1.5 Eine differenzierte Betrachtung nach Erfahrung mit VERA8

Da in dieser Arbeit die Bearbeitung von Tests auch als von einer Testkompetenz abhängig angenommen wird und man diese Testkompetenz als Lehrkraft auch erst erlernen muss, um sie seinen Schülern vermitteln zu können, spielt möglicherweise auch die Erfahrung mit zentralen Lernstandserhebungen bei der Vorbereitung auf VERA8 eine Rolle. Die Lehrkräfte wurden daher gebeten anzugeben, ob sie zum ersten Mal eine Klasse unterrichteten, die an VERA8 teilnahm, oder ob sie bereits in Schuljahren davor Erfahrungen mit VERA8 gesammelt haben. In welchem Umfang diese Erfahrung bestand, wurde nicht differenzierter erhoben. Aber die Lehrkräfte mit Erfahrung sollten zusätzlich angeben, ob sie ihrer Einschätzung nach in diesem Schuljahr intensiver, genauso intensiv oder weniger intensiv als beim allerersten Mal vorbereiteten. Einige der bisherigen Analysen sollen daher nun noch einmal etwas differenzierter vorgenommen werden. Dabei wird auf eine Unterscheidung zwischen den vier Teilstudien an dieser Stelle verzichtet, da die Fallzahlen sonst zu gering sind und sich außer der für die Vorbereitung aufgewendeten Unterrichtszeit keine sinnvoll erklärbaren Unterschiede zwischen den Teilstudien zeigten. Für die Unterrichtszeit ist dies in Tab. 7.8 auch für die vier Teilstudien ausgewiesen, wird aber nicht ausführlich betrachtet, da die Tendenz für alle vier Teilstudien mit einer Ausnahmen[126] auch für die Vorbereitungszeit mit dem globalen Ergebnis übereinstimmt.

[126] Für die Teilstudie A2 zeigt sich ein höherer Mittelwert der aufgewendeten Unterrichtszeit für Lehrkräfte, die glaubten gleich bleibend vorzubereiten, im Vergleich zu Lehrkräften, die ihre Vorbereitung als intensiver

Unterscheidet man nur zwischen Lehrkräften, die erstmalig mit VERA8 konfrontiert sind, und solchen, die bereits Erfahrungen mit VERA8 gesammelt haben, zeigt der Welch-Test[127] einen signifikanten Unterschied zwischen beiden Gruppen bzgl. der aufgewendeten Vorbereitungszeit (t[559.53]=2.150, p=0.03), wobei die erste Gruppe durchschnittlich fast eine Unterrichtsstunde weniger für die Vorbereitung aufwendet (M=6.89, SE=.31, SD=4.71) als die zweite Gruppe (M=7.77, SE=.27, SD=5.61). Exkludiert man zusätzlich aus der zweiten Gruppe diejenigen Lehrkräfte, die angeben, weniger oder mehr als zuvor für die Vorbereitung aufzuwenden, so beträgt der Unterschied nun über eine Unterrichtsstunde (Gruppe 2_b: M=8.35, SE=.31, SD=6.13) und der Unterschied ist auch signifikant auf Niveau p=.00 (t[493.82]=3.024) mit allerdings niedriger (normierter) Effektstärke (r=.13). Diejenigen Lehrkräfte aus der Betrachtung herauszunehmen, die die für die Vorbereitung aufgewendete Unterrichtszeit reduziert haben, erscheint sinnvoll und wird daher auch in den weiteren Analysen derart gehandhabt.

Als weitere Intensitätsindikatoren können wiederum die Nutzungsvariabilität als Summenindex der FA-Maßnahmen (a, b, c, d, e, h) und das Ausmaß der Lehrkräfte betrachtet werden, die FA- oder CA-Maßnahmen durchführten (vgl. Abb. 7.9). Im Gegensatz zum zeitlichen Umfang der Vorbereitung zeigt sich für den Summenindex der FA-Maßnahmen kein deutlicher Unterschied zwischen den beiden Gruppen (Lehrkräfte ohne VERA8-Erfahrung, Lehrkräfte mit VERA8). Zwar gibt es in Gruppe 2_b leicht mehr Lehrkräfte die vier oder mehr Maßnahmen durchgeführt haben, aber der Unterschied ist mit p=.117 nicht signifikant (χ^2[7]=10.206).

Tabelle 7.7

Vorbereitungszeit nach Erfahrung und eingeschätzter Veränderung der Intensität – insgesamt

Erfahrung	aggregierte Daten		
	n	M(SD)	95% CI
bisher keine Erfahrung mit VERA8	237	6.89(4.71)	[6.29, 7.46]
Intensität gleichbleibend	268	8.35(6.13)	[7.59, 9.17]
Intensivere Vorbereitung	61	9.52(4.66)	[8.48, 10.79]
Vorbereitung weniger intensiv	106	5.27(3.65)	[4.68, 6.02]

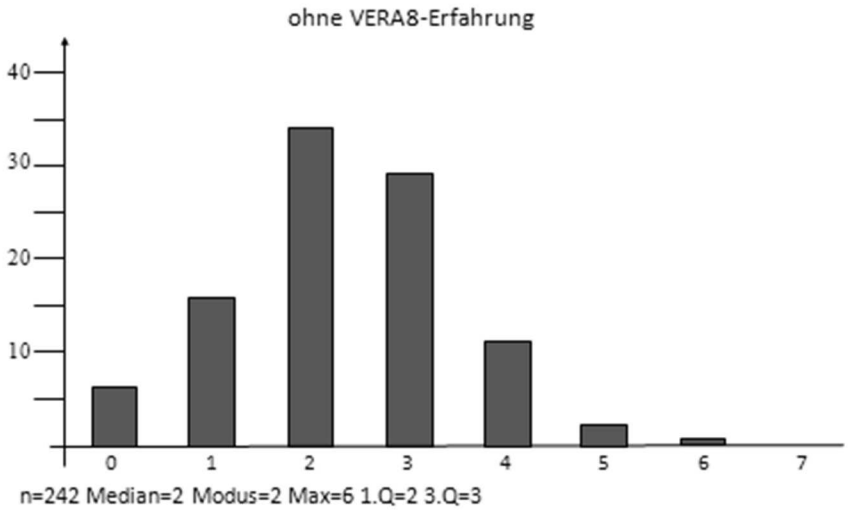
einschätzen. Der Unterschied resultiert wahrscheinlich aus der statistischen Ungenauigkeit, die sich aus der geringen Fallzahl für Lehrkräfte mit intensiverer Vorbereitung ergibt.

[127] H0: die Varianzen sind gleich muss bei p=.05 abgelehnt werden. Daher wird wie oben erläutert kein t-Test nach Student durchgeführt.

Tabelle 7.8

Vorbereitungszeit nach Erfahrung und eingeschätzter Veränderung der Intensität – differenziert nach Teilstudien

Erfahrung	Studie A1			Studie A2			Studie B1			Studie B2		
	n	M(SD)	95% CI	N	M(SD)	95% CI	n	M(SD)	95% CI	n	M(SD)	95% CI
bisher keine Erfahrung mit VERA8	72	6.33(4.82)	[5.28, 7.49]	56	7.11(5.07)	[5.82, 8.38]	73	6.79(4.35)	[5.85, 7.77]	36	7.86(4.67)	[6.39, 9.33]
Intensität gleichbleibend	70	7.00(5.57)	[6.30, 12.50]	77	9.95(7.63)	[8.27, 11.67]	80	8.03(5.50)	[6.86, 9.31]	41	8.30(4.27)	[7.03, 9.56]
Intensivere Vorbereitung	10	9.50(5.34)	[5.86, 8.39]	23	8.78(3.84)	[7.22, 10.43]	20	8.75(3.34)	[7.35, 10.25]	8	13.63(7.07)	[9.25, 18.50]
Vorbereitung weniger intensiv	36	4.69(3.25)	[3.72, 5.69]	26	6.58(3.62)	[51.9, 8.04]	32	5.00(3.33)	[3.94, 6.06]	12	4.92(5.21)	[2.67, 8.17]



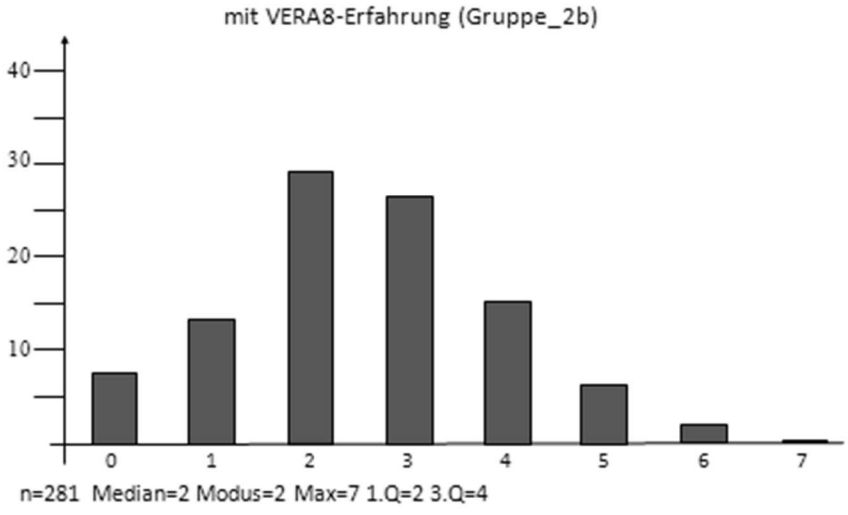


Abbildung 7.9: Darstellung der Nutzungsvariabilität von FA-Maßnahmen für Lehrkräfte mit und ohne VERA8-Erfahrung (Verteilung der Lehrkräfte in Prozent)

Entsprechend verhält es sich mit der Anzahl an Lehrkräften, die zumindest eine Maßnahme im Sinne eines FA durchführten. Hier sind die Werte mit 93.0% (Gruppe 1) und 93.0% (Gruppe 2_b) identisch. Anders stellt sich die Lage bei den CA-Maßnahmen dar. Der Anteil der Lehrkräfte, die als Vorbereitung auf VERA8 Wiederholungs- und Übungsphasen durchführten, ist für die erfahrenen Lehrkräfte (Gruppe 2_b) um über zehn Prozent höher (83.3%) als für die Lehrkräfte ohne Erfahrung mit VERA8 (74.4%) und der Unterschied ist signifikant ($\chi^2[1]=6.234$) mit p=.013, weist allerdings mit (normierter) Effektstärke w=.11 wiederum nur einen niedrigen Effekt auf. Dabei resultiert der Unterschied vor allem aus dem Befunde, dass Lehrkräfte mit VERA8-Erfahrung häufiger eine Wiederholung der Inhaltsbereiche (47.0% zu 38.0%) und (aller oder einiger) Prozesskompetenten selbst im (33.5% zu 26.0%) Unterricht durchführen.

Tabelle 7.9

Anteil der Lehrkräfte, die Familiarity-Approach- oder Content-Approach -Maßnahmen durchführten

	keine Erfahrung mit VERA8 (n=242)	Erfahrung mit VERA8 (n=281)
Maßnahmenarten	absolut (%)	Absolut (%)
Lehrkräfte, die mindestens eine der vorformulierten oder freiformulierten FA-Maßnahmen umsetzten	225 (93.0%)	323 (92.9%)
Lehrkräfte, die mindestens eine der vorformulierten oder freiformulierten CA-Maßnahmen umsetzten	180 (74.4%)	292 (83.3%)

Abschließend soll im Rahmen dieser Forschungsfrage noch einmal ein analysiert werden, ob die drei aufgabenbezogenen Themen (q), (p) und (r) in beiden Gruppen unterschiedlich häufig angesprochen wurden. Wie bereits erläutert wurde, setzen diese eine Reflexion der Lehrkraft über die Funktionsweise von VERA8-Aufgaben voraus und insbesondere (q) stellt dabei eine Herausforderung dar. Möglich ist daher, dass Lehrkräfte ohne Erfahrung mit VERA8 sich häufiger diese Reflexion nicht zutrauen bzw. sie nicht vollzogen haben und daher die drei Themen weniger häufig ansprechen.

Die Ergebnisse sind in Tab. 7.10 dargestellt: Für alle drei Themen zeigt sich, dass erfahrene Lehrkräfte die drei Themen häufiger mehrfach angesprochen haben (zwischen 61.4% bei Thema (q) und 63.4% bei Thema (r)) als die Kollegen ohne Erfahrung mit VERA8 (bei (q) 45.2%, bei (r) 46.5%). Auch sprachen unerfahrene Lehrkräfte die drei Themen häufiger gar nicht an. Die Unterschiede sind alle auf Niveau p=0.001 signifikant ($\chi^2_p[2]=19.468$, w=.20; $\chi^2_q[2]=13.484$, w=.16; $\chi^2_r[2]=14.338$, w=.17), aber erneut nur mit geringer Effektstärke.

Tabelle 7.10

Welche der folgenden Themen haben Sie im Unterricht angesprochen? – Differenzierung nach Erfahrung mit VERA8

	keine Erfahrung mit VERA8 (n=226)			Erfahrung mit VERA8 (n=272)		
Maßnahme	mehrfach (%)	einmal (%)	gar nicht (%)	mehrfach (%)	Einmal (%)	gar nicht (%)
(p) wie man die Aufgabenstellung der LSE-Aufgaben richtig versteht	107 (46.5%)	70 (28.9%)	53 (21.9%)	178 (63.3%)	45 (16.0%)	51 (18.1%)
(q) wie man den Aufgabenstellungen der LSE-Aufgaben die wichtigen Informationen entnimmt	103 (45.2%)	56 (24.6%)	69 (30.3%)	167 (61.4%)	43 (15.8%)	62 (22.8%)
(r) wie man in Hinblick auf die speziellen Antwortformate richtig antwortet	105 (46.5%)	67 (29.6%)	54 (23.9%)	173 (63.4%)	56 (20.4%)	44 (16.1%)

7.1.6 Schwerpunktsetzungen mit und ohne Blick auf VERA8 für das Erhebungsschuljahr nach Einschätzung der Lehrkräfte

Schwerpunktsetzungen werden hier auf Lehrerebene als im Vergleich zu den vorherigen Schuljahren intensivere oder weniger intensivere Behandlung einer Prozesskompetenz verstanden. Dabei kann diese Verschiebung durch Überlegungen im Zusammenhang mit VERA8 begründet sein, wobei dies aus den Ergebnissen der vergangenen Jahren resultieren können oder auch aus der Antizipation der bevorstehenden Testaufgaben. Dies wurde in den Teilstudien nicht weiter differenziert, Lehrkräfte ohne Erfahrung mit VERA8 wurden allerdings aus logischen Gründen nicht bei diesem Analyseschritt berücksichtigt. Auf

Fachgruppenebene kann es zu Absprachen über ein intensives oder weniger intensives Unterrichten einzelner Bereiche kommen. Hier wurde nicht explizit nach Prozesskompetenzen gefragt, sondern allgemein nach einzelnen Bereichen, die in einer anderen Intensität unterrichtet werden sollten als zuvor. Auch dabei können wiederum Überlegungen zu VERA8 als Begründung dienen.

Tabelle 7.11

Veränderungen in der Gewichtung einzelner (Prozess-)Bereiche durch einzelne Lehrkraft im Zusammenhang mit VERA8 – differenziert nach Teilstudien

	Studie A1 (n=117)	Studie A2 (n=131)	Studie B1 (n=137)	Studie B2 (n=62)
Gewichtung	absolut (%)	absolut (%)	absolut (%)	Absolut (%)
Prozesskompetenz intensiver	28 (23.9%)	36 (27.5%)	33 (24.1%)	19 (30.6%)
davon wegen VERA8	9	18	7	3
weniger intensiv	5 (4.3%)	12 (9.2%)	9 (6.6%)	8 (12.9%)
davon wegen VERA8	1	3	3	3

Auf Ebene der einzelnen Lehrkraft gaben zwischen 23.9% (A1) und 30.6% (B2) der Lehrkräfte an, ihren Unterricht zu Schuljahresbeginn derart geändert zu haben, dass sie im aktuellen Schuljahr der Befragung (2009/10) eine Prozesskompetenz intensiver behandelt haben als in den Jahren davor. Von diesen Lehrkräften führten aber durchschnittlich über alle vier Teilstudien nur gut dreißig Prozent die Veränderung auf Überlegungen zurück, die mit VERA8 im Zusammenhang stehen. Umgekehrt liegt die Zahl der Lehrkräfte, die eine Prozesskompetenz nach eigenen Angaben weniger intensiv unterrichtet hat, jeweils deutlich niedriger zwischen 12.9% (B2) und 4.3% (A1). Von diesen Lehrkräften sahen wiederum nur zehn Lehrkräfte insgesamt einen Zusammenhang zu VERA8.

Tabelle 7.12

Veränderungen in der Gewichtung einzelner (Prozess-)Bereiche durch die Fachgruppe im Zusammenhang mit VERA8 – differenziert nach Teilstudien

	Studie A1 (n=181)	Studie A2 (n=185)	Studie B1 (n=198)	Studie B2 (n=99)
Gewichtung	absolut %	absolut %	absolut %	absolut %
Bereich in FK intensiver	21 (11.6%)	29 (15.7%)	23 (11.6%)	13 (13.1%)
davon wegen VERA8	12	16	12	5
Bereich weniger intensiv in FK	4 (2.2%)	6 (3.2%)	3 (1.5%)	5 (5.1%)
davon wegen VERA8	1	3	3	1

Im niedrigen zweistelligen Bereich liegen auch die Anzahl der Fachgruppen, die gemeinsam beschlossen haben, eine Prozesskompetenz intensiver zu behandeln. Hiervon berichteten zwischen 15.7% (A2) und 11.6% (A1 & B1). Die Hälfte dieser Lehrkräfte führte dies auf Überlegungen zu VERA8 zurück. Umgekehrte Überlegungen von Fachgruppen wurden noch seltener berichtet, als dies für Lehrerebene angegeben wurde.

7.1.7 Nutzung von Vorbereitungs- und Kompetenzheften

Zur Vorbereitung auf VERA8 setzten Lehrkräfte mehrheitlich Vorbereitungshefte oder Kompetenzhefte ein bzw. empfahlen ihren Schülern die Anschaffung dieser Hefte (s. Tab. 7.13 und Tab. 14). Während Kompetenzhefte eindeutig dazu dienen, Inhalts- und Prozessbereiche zu wiederholen, ohne dass die Gestaltung der Aufgaben eine wesentliche Rolle spielt, zeichnen sich Vorbereitungshefte gerade dadurch aus, dass die Übungs- oder Wiederholungsphasen mit Aufgaben durchgeführt werden können, die zumindest in Teilen den VERA8-Testaufgaben nachempfunden sein sollen. Kompetenzhefte sind daher nur im Sinne eines Content Approach zu nutzen, Vorbereitungshefte können hingegen auch einem Familiarity Approach dienen.

Tabelle 7.13

Einsatz von Vorbereitungsheften im Unterricht - differenziert nach Teilstudien

Maßnahme	Studie A1 (n=191) absolut %	Studie A2 (n=189) absolut %	Studie B1 (n=212) absolut %	Studie B2 (n=103) Absolut %
Vorbereitungshefte eingesetzt	100 (52.4%)	107 (56.6%)	121 (57.1%)	59 (57.3%)
ausschließlich Vorbereitungshefte eingesetzt	82 (42.9%)	77 (40.7%)	85 (40.1%)	47 (45.6%)

Tabelle 7.14

Einsatz von Kompetenzheften im Unterricht - differenziert nach Teilstudien

Maßnahme	Studie A1 (n=175) absolut %	Studie A2 (n=182) absolut %	Studie B1 (n=202) absolut %	Studie B2 (n=92) Absolut %
Kompetenzhefte eingesetzt	34 (19.4%)	43 (23.6%)	48 (23.8%)	17 (18.5%)
ausschließlich Kompetenzhefte eingesetzt	16 (9.1%)	13 (7.1%)	12 (5.9%)	5 (5.4%)
beide Arten eingesetzt	18 (10.3%)	30 (16.5%)	36 (17.8%)	12 (13.0%)

Es zeigt sich über alle vier Teilstudien, dass mindestens eine der beiden Heftarten von ca. sechzig Prozent der Lehrkräfte (A1: 60.7%, A2: 63.5%, B1: 62.7%, B2: 62.1%) im Unterricht eingesetzt wurde. Dabei griffen die Lehrkräfte wesentlich häufiger ausschließlich auf Vorbereitungshefte zurück oder setzen beide Heftarten ein (in den Teilstudien jeweils häufiger als der ausschließliche Einsatz von Kompetenzheften). Auch dieses Bild weist nur minimale Differenzen zwischen den vier Teilstudien auf. Somit kann gefolgert werden, dass diesen Lehrkräften ein FA durchaus wichtig war und sie diese zusätzlichen Materialien auch

einsetzten, um die Schüler mit den Besonderheiten der Tests vertraut zu machen. Gleichwohl darf dies nicht damit verwechselt werden, dass ein CA keine Rolle spielte. Auch Vorbereitungshefte dienen vorwiegend der Übung und Wiederholung von Inhaltsbereichen bzw. Leitideen und (etwas untergeordnet) Prozessbereichen bzw. Kompetenzen. Die Bedeutung der inhaltlichen Vorbereitung ergibt sich auch aus den 10.3% (A1) bis 17.8% (B1) der Lehrkräfte, die beide Heftarten einsetzten.

Lehrkräfte, die Vorbereitungs- oder Kompetenzhefte oder sogar beide Heftarten einsetzten, wurden gebeten, zusätzlich anzugeben, ob sie damit Übungsphasen für Inhaltsbereiche oder Prozesskompetenzen durchführten und ob alle Inhalts- und alle Prozessbereiche wiederholt wurden oder nur einige ausgewählte. Die Ergebnisse sind in den beiden folgenden Tabellen dargestellt (Tab. 7.15 und Tab. 16).

Tabelle 7.15

Wiederholung von Inhalts- und Prozessbereiche mit Vorbereitungsheften – differenziert in vier Teilstudien

	Studie A1	Studie A2	Studie B1	Studie B2
	(n=99)	(n=115)	(n=119)	(n=61)
	absolut	Absolut	absolut	Absolut
Nutzung von Vorbereitungsheften	%	%	%	%
alle Inhaltsbereiche wiederholt	26 (26.3%)	35 (30.4%)	40 (33.6%)	9 (14.8%)
einzelne Inhaltsbereiche wiederholt	70 (70.7%)	72 (62.6%)	77 (64.7%)	51 (83.6%)
	(n=98)	(n=113)	(n=119)	(n=59)
alle Prozessbereiche wiederholt	17 (17.3%)	26 (23.0%)	30 (25.2%)	8 (13.6%)
einzelne Prozessbereiche wiederholt	73 (74.5%)	74 (65.5%)	81 (68.1%)	44 (74.7%)

Anmerkung: Die Anteile ergänzen sich zu 100% mit denjenigen, die „weder noch“ angaben.

Auf den ersten Blick zeigt sich für die Vorbereitungshefte kein deutlicher Unterschied zwischen den Wiederholungsphasen für Inhaltsbereiche und Prozessbereiche. Lehrkräfte

gaben an, die Vorbereitungshefte mehrheitlich genutzt zu haben, um nur einige Bereiche zu wiederholen und üben zu lassen. Nur maximal ein Drittel (B1) der Lehrkräfte hat mit den Heften eine Wiederholungsphase für alle Inhaltsbereiche durchgeführt. Für die Prozessbereiche liegt der Anteil der Lehrkräfte, die alle vier Prozessbereiche wiederholten bei maximal einem Viertel (wiederum B1) und damit leicht unter dem Wert für die Inhaltsbereiche. Entsprechend drückt sich dieser kleine Unterschied auch darin aus, dass keine Lehrkraft, die alle Prozessbereiche wiederholte, nicht auch zumindest einige Inhaltsbereiche wiederholten wollte und nur fünf der Lehrkräfte, die mit den Heften überhaupt keine Inhaltsbereiche wiederholten, trotzdem zumindest einige Prozessbereiche üben ließen. Andere Zwecke außer Übungsphasen verfolgten Lehrkräfte nur vereinzelt, über alle vier Teilstudien gaben nur neun Lehrkräfte an, die Hefte eingesetzt zu haben und mit ihnen weder Inhalts- noch Prozessbereiche wiederholt zu haben.

Tabelle 7.16

Wiederholung von Inhalts- und Prozessbereiche mit Kompetenzheften – differenziert in vier Teilstudien

	Studie A1 (n=45)	Studie A2 (n=48)	Studie B1 (n=53)	Studie B2 (n=21)
	absolut	Absolut	absolut	Absolut
Nutzung von Kompetenzheften	%	%	%	%
Kompetenzhefte *alle* Inhaltsbereiche wiederholt	5 (11.1%)	10 (20.8%)	7 (13.2%)	0
Kompetenzhefte *einzelne* Inhaltsbereiche wiederholt	30 (66.7%)	30 (62.5%)	43 (81.1%)	19 (90.5%)
	(n=43)	(n=48)	(n=53)	(n=20)
alle Prozessbereiche wiederholt	6 (14.0%)	7 (14.6%)	7 (13.2%)	0
einzelne Prozessbereiche wiederholt	25 (58.1%)	29 (60.4%)	40 (75.5%)	18 (90.0%)

Anmerkung: Die Anteile ergänzen sich zu 100% mit denjenigen, die „weder noch“ angaben.

Für die Kompetenzhefte sieht das Bild ähnlich aus. Auch hier wurden die Hefte vorwiegend genutzt, um bestimmte Bereiche zu wiederholen und Inhaltsbereiche wurden nur minimal

umfassender wiederholt als Prozessbereiche, wenn Kompetenzhefte eingesetzt wurden. Insgesamt gaben auch nur zwei Lehrkräfte an, alle Prozessbereiche, aber nur einige Inhaltsbereiche mit diesen Heften wiederholt zu haben.

7.1.8 Außerunterrichtliche Vorbereitung

Vorbereitungs- und Kompetenzhefte können nicht nur für Wiederholungs- und Übungsphasen während der Unterrichtszeit dienen, sondern auch für die außerunterrichtliche Vorbereitung genutzt werden. Eine Vorbereitung findet dann nicht zwingend innerhalb der Unterrichtszeit statt, kann aber trotzdem von den Lehrkräften initiiert worden sein. Es wurde daher gezielt gefragt, ob Lehrkräfte ihren Schülern die Anschaffung von Vorbereitungs- oder Kompetenzheften nahe gelegt haben.

Selbstverständlich kann eine Vorbereitung außerhalb des Unterrichts auch ohne Vorbereitungs- oder Kompetenzhefte geschehen. Weiterhin wurde gefragt, ob Lehrkräfte einzelnen Schülern oder sogar der ganzen Klasse empfohlen haben, sich gezielt auf VERA8 vorzubereiten. Möglich ist auch, dass Nachhilfeangebote von den Schülern in entsprechende Richtung in Anspruch genommen werden oder Lehrkräfte dies empfehlen. Schließlich wurde erhoben, ob Lehrkräfte bemerkt haben, dass ihre Schüler im außerunterrichtlichen Bereich besonders für VERA8 geübt haben.

Tabelle 7.17

außerunterrichtliche Übungsphasen in Wahrnehmung der Lehrkräfte – differenziert nach Teilstudien

	Studie A1		Studie A2		Studie B1		Studie B2	
	Ja	Nein	Ja	Nein	Ja	Nein	Ja	Nein
	absolut	Absolut	absolut	absolut	absolut	absolut	absolut	Absolut
Maßnahme	%	%	%	%	%	%	%	%
Haben Sie einzelnen Schülern empfohlen, sich speziell auf VERA8 vorzubereiten?	7 (3.8%)*	179 (96.2%)	4 (2.1%)*	185 (97.9%)	8 (3.9%)*	196 (96.1%)	7 (6.7%)*	95 (93.3%)

	Studie A1		Studie A2		Studie B1		Studie B2	
	Ja	Nein	Ja	Nein	Ja	Nein	Ja	Nein
	absolut	Absolut	absolut	absolut	absolut	absolut	absolut	Absolut
Maßnahme	%	%	%	%	%	%	%	%
Haben Sie der Klasse empfohlen, sich speziell auf VERA8 vorzubereiten?	103 (55.1%)	78 (44.9%)	132 (70.6%)	55 (29.4%)	130 (64.7%)	71 (35.3%)	65 (63.1%)	38 (36.9)
Nehmen einzelne Schüler Nachhilfe speziell mit Blick auf VERA8?	12 (6.4%)	176 (93.6%)	7 (3.7%)	181 (96.3%)	10 (5.1%)	188 (94.9%)	7 (7.0%)	93 (93.0%)
Haben die Schüler für VERA8 außerhalb des Unterrichts besonders geübt?	103 (57.9%)	75 (42.1%)	103 (57.9%)	75 (42.1%)	114 (60.6%)	74 (39.4%)	49 (55.1%)	40 (44.9%)

Anmerkung: *Es wurden nur diejenigen Lehrkräfte berücksichtigt, die nicht gleichzeitig der ganzen Klasse empfohlen haben, sich auf VERA8 vorzubereiten.

Das Bild der außerunterrichtlichen Vorbereitung zeigt sich über alle vier Teilstudien als ein eher homogenes. Nur ein niedriger einstelliger Anteil der Lehrkräfte hat gezielt einzelnen Schülern geraten, sich speziell auf VERA8 vorzubereiten. Auch nur wenige Lehrkräfte haben ihren Schülern (einigen einzelnen oder der gesamten Klasse) die Anschaffung von Vorbereitungsheften (insgesamt zwölf Lehrkräfte) oder Kompetenzheften (insgesamt elf Lehrkräfte) empfohlen, ohne diese selbst im Unterricht einzusetzen. Mehrheitlich (zwischen 55.1% [A1] und 70.6% [A2]) wurde aber der gesamten Klasse eine Vorbereitung nahe gelegt.

Trotzdem zeigt sich auch ein relativ großer Anteil von Lehrkräften, die angaben, ihrer Klasse gar keine Vorbereitung empfohlen zu haben. Dieser liegt in allen vier Teilstichproben bei über einem Viertel der Befragten, obwohl nicht einmal sieben Prozent der Lehrkräfte keine Vorbereitung im Unterricht durchgeführt haben. Denjenigen Lehrkräften, die zwar eine Vorbereitung im Unterricht durchgeführt haben, aber ihren Schüler dies selbst nicht empfohlen haben, sind die einzelnen Vorbereitungsmaßnahmen entweder nicht als solche bewusst oder sie sahen die Vorbereitung nicht als Aufgabe der Schüler an bzw. trauten ihnen dieses nicht zu. In dieser Richtung stellen sich auch die weiteren Befunde zur außerunterrichtlichen Vorbereitung dar. Auffällig ist dabei in den Teilstudien A2, B1 und B2 der Anteil der Lehrkräfte, die ihrer Klasse zwar eine Vorbereitung empfohlen haben, aber nicht wahrgenommen haben, dass sich ihre Schüler tatsächlich außerhalb des Unterrichts

auf VERA8 vorbereitet haben. Nur in Teilstudie A1 gaben tatsächlich gleichviele Lehrkräfte an, ihrer Klasse eine außerunterrichtliche Vorbereitung empfohlen zu haben und diese bei ihren Schülern dann auch tatsächlich wahrgenommen zu haben. Auch befand sich der Anteil der Klassen, in denen (nach Kenntnis der Lehrkraft) einzelne Schüler speziell mit Blick auf VERA8 Nachhilfe genommen haben, ebenfalls nur im einstelligen Bereich. Dies kann sich aber auch dadurch erklären, dass Lehrkräften weder bekannt sein muss, ob ihre Schüler Nachhilfe nehmen, noch, aus welcher Intention dies geschieht. Explizit empfohlen hat insgesamt auch nur eine einzige Lehrkraft Schülern, im Vorfeld von VERA8 speziell Nachhilfe in Anspruch zu nehmen.

Gleichwohl nahmen in allen vier Teilstudien mehr als die Hälfte der Lehrkräfte wahr, dass ihre Schüler für VERA8 besonders geübt haben. Dabei wurde zwar nicht unterschieden, ob dies auf alle oder nur einen Teil der Klasse zutraf, es verdeutlicht aber das Bewusstsein der Schüler für VERA8. Damit zeigt sich die Vorbereitung außerhalb des Unterrichts als weniger häufig durch Lehrkräfte initiiert und beobachtet, aber trotzdem als Größe von bedeutsamem Umfang. Dies wird insbesondere darin deutlich, dass Lehrkräfte, wenn sie eine Vorbereitung nahe legten, dies nicht individuell machten, sondern ihren Klassen global empfahlen.

Nachdem nun ein erster Eindruck über die Datenlage gewonnen werden konnte, dient der folgende Abschnitt dazu, Modelle zu gewinnen, auf deren Grundlage die deskriptiven Daten einer tieferen Analyse unterzogen werden können.

7.2 Ergebnisse der Modellvergleiche

Im folgenden Abschnitt werden die fünf Modelle mit ihren Kennzahlen präsentiert und inhaltlich interpretiert. Mittels dieser Kennzahlen und den gegebenen Interpretationen sollen die Klassenlösungen und anschließend die Modelle innerhalb der jeweiligen Modellgruppe einander gegenübergestellt werden, um für jede Gruppe jeweils das beste Modell für die weitergehende Analyse des Vorbereitungsverhaltens zu bestimmen. Als Resultat dieses Abschnitts sollen sich folglich zwei Modelle herauskristallisieren, die für eine detailliertere Analyse des Vorbereitungsverhaltens genutzt werden sollen. Die drei Modelle M11, M11a und M12 repräsentieren den Ansatz, Teildomänen der Lehrerhandlungskompetenz als Prädiktoren des Vorbereitungsverhaltens anzunehmen. Die speziellen LCA wurden für diese Modelle genau wie für Modell M21 auf Grundlage des Datensatzes aus Studie A geschätzt. Die Modell M21 und M22 folgen hingegen der Vorstellung, dass sich das Vorbereitungsverhalten der Lehrkräfte durch Prädiktoren erklären lässt, die nach Erkenntnissen der „Feedbackforschung“ dazu beitragen, wie Personen mit Feedback umgehen. Im Unterschied zu M21 werden bei Modell M22 drei Skalen (zur Rezeption von VERA8-Ergebnissen, zu ihrer Reflexion und zu evtl. abgeleiteten

Veränderungen) einbezogen, die einen direkten inhaltlichen Bezug auf VERA8 (bzw. LSE8) nehmen. Darüber hinaus basiert das Modell M22 als einziges Modell auf dem Datensatz der Studie B.

7.2.1 Modelle mit personenbezogene Überzeugungen und Überzeugungen über das Lehren & Lernen als Prädiktoren des Vorbereitungsverhaltens

Modelle mit ausschließlich personenbezogenen Überzeugungen

Das Modell M11 besteht, wie in 6.4 beschrieben wurde, aus Skalen zu personenbezogenen Überzeugungen, nämlich dem beruflichen Beanspruchungserleben (BEL), dem Overcommitment (OC), dem Arbeitsengagement (UWES), dem fachdidaktischen Fähigkeitsselbstkonzept (SK) und der Selbstwirksamkeit (SWE). Für alle fünf Skalen wie auch für die Gewissenhaftigkeit (GW) legen die Kennzahlen nahe, eine Vier-Klassen-Lösung für die einzelnen Skalen-Items zu verwenden. Entsprechend werden keine Klassenlösungen dargestellt, die auf dichotomen Skalen-Items beruhen.

Die bewusst gewählte Analogie des Modells zur Typen-Klassifikation des Arbeitsbezogenes Verhaltens- und Erlebensmuster (AVEM) legt auch für die Modellklasseneinteilung eine Vier-Klassen-Lösung nahe. Es wurden trotzdem auch latent Klasseneinteilungen mit Drei-Klassenlösungen und Fünf-Klassenlösungen geschätzt. Die entsprechenden Kennzahlen sind in Tabelle 7.18 ausgewiesen. Dort sind für jede Klassenlösung die zu schätzenden Parameter, der Likelihood-Wert (L), die AIC- und BIC-Werte, die Werte für den Likelihood-Quotiententest (LQT) und den Pearson'schen χ^2-Test (CHI) sowie die durchschnittliche Zuordnungswahrscheinlichkeit der Lehrkräfte zu der ihnen jeweils zugeordneten Klasse angegeben. Der Tabelle kann man entnehmen, dass die Kennzahlen der Informationskriterien AIC und BIC widersprüchliche Tendenzen aufweisen, welche Klassenlösung die wahrscheinlich Beste ist. Mit fünf Items ist die Itemzahl zwar gering, sodass eher das AIC berücksichtigt werden sollte, gleichzeitig entspricht die Zahl der beobachteten 372 Fälle aber bei weitem nicht den 1024 möglichen Mustern. Jener Überlegung nach, sollte der BIC-Wert höher gewichtet werden. Den kleinsten AIC-Wert weist die Fünf-Klassenlösung auf, gefolgt von der Vier-Klassenlösung. Umgekehrt besitzt die Drei-Klassenlösung den geringsten BIC-Wert, wiederum gefolgt von der Vier-Klassenlösung. Ungeachtet dessen sind die Pearson'schen Werte alle signifikant höher als es für die jeweiligen durch Bootstrapping erzeugten χ^2-Verteilungen akzeptablen wären. Gleiches gilt im Besonderen noch deutlicher für die Werte zum Likelihood-Quotienten-Test. Für keine der drei Lösungen kann dementsprechend angenommen werden, dass sie zu den Daten passt. Die drei Klassenlösungen wurden demzufolge nicht für die weitere Analyse des Vorbereitungsverhaltens herangezogen.

Ungeachtet der unzureichenden statistischen Passung der drei Klassenlösungen lassen sich aber die geschätzten Klassenlösungen interpretieren. Da mit dem AVEM eine Einteilung in vier Typen als Referenz benutzt wurde, um die Skalen des Modells M11 auszuwählen, sollen die Klassen der 4-Klassenlösung daher trotzdem interpretiert werden. Betrachtet man die drei linken Skalen (BEL, OC, UWES), erkennt man zu den vier Typen des AVEM durchaus ähnliche Klassen.

Tabelle 7.18

Kennzahlen zum Modell M11 mit vierstufigen Skalen-Items

Klassenzahl	Parameter	L	AIC	BIC	LQT	CHI	Zuordnungs-wahrschein-lichkeit
3	47	-2221.76	4537.511	4721.825	694.2513	1186.166	85.5%
4	63	-2193.10	4512.198	4758.257	635.8461	1027.827	86.0%
5	79	-2175.37	4508.748	4818.552	600.3189	978.839	86.5%

Es ergibt sich eine Klasse (als Typ A bezeichnet), die die schlechtesten Werte im Bereich der erlebten Beanspruchung aufweist und auch ein starkes Overcommitment zeigt, gleichzeitig gute Werte im Arbeitsengagement besitzt. Im Vergleich zum originalen Typ A des AVEM zeigen die hier klassifizierten Lehrkräfte aber nicht das größte Arbeitsengagement. Zusätzlich besitzen Lehrkräfte dieses Typs das positivste Fähigkeitsselbstkonzept und die größte Selbstwirksamkeitserwartung. Der Klasse wurden 18.0% der klassifizierten Lehrkräfte zugeordnet. Dies sind weniger als die ca 25%, die Schaarschmidt dem originalen Typ A zuweist. Ähnliche Werte wie Typ A zeigt in den Bereichen BEL und OC auch das als Typ B bezeichnete Muster. Allerdings zeigen Lehrkräfte dieses Typs ein viel geringeres Arbeitsengagement und verfügen zusätzlich über ein negatives Fähigkeitsselbstkonzept und eine sehr geringe Selbstwirksamkeitserwartung. 22.8% der Lehrkräfte wurden dieser Klasse zugeordnet. Auch die originale Klassegröße des Typs B wird damit unterboten.

Konträr zu Typ A ist das als Typ S bezeichnete Muster, zu dem Lehrkräfte gehören, die sehr gute Werte im BEL und gute Werte im OC aufweisen und ein mittleres Arbeitsengagement zeigen. Ihr Fähigkeitsselbstkonzept und ihre Selbstwirksamkeitserwartung sind ebenfalls im mittleren Bereich. Hier ist einer der deutlichsten Unterschiede zu originalen Klassifikation von Schaarschmidt und Kollegen sichtbar, denn nach der ursprünglichen Klassifikation zeigen dem Typ S zugeordnete Personen das geringste Engagement. In der Vier-Klassenlösung des Modells M11 sind es hingegen Lehrkräfte, die dem Typ B zugeordnet wurden. Das Muster

Typ S ist in der Studie für 41.3% der Lehrkräfte als die wahrscheinliste Klassenzugehörigkeit geschätzt worden. Die Klasse ist damit noch einmal deutlich größer als in der von Schaarschmidt berichteten Einteilung. Dort wurde ca. ein Drittel dem Typ S zugeordnet. Gute Werte im Fähigkeitsselbstkonzept und ansonsten sehr gute Werte zeichnen das Muster Typ G aus. Für die linken Skalen zeigen Lehrkräfte dieses Musters jeweils die besten Werte. Auch in dieser Klasse befinden sich 18.0% der Lehrkräfte, während es in der originalen Verteilung ca. ein Viertel ist.

Zusammenfassend lässt sich für die Vier-Klassenlösung des Modells M11 festhalten, dass sich in der Tat die erwartete Typenstruktur in groben Zügen abzeichnet. Gleichzeitig gibt es aber Unterschiede in der Auftrittshäufigkeit der vier Klassen und mehr oder weniger starke Abweichungen in einzelnen Bereichen. Besonders auffällig ist der Unterschied beim Muster Typ A bezüglich des Arbeitsengagements. Dabei muss aber noch einmal klar unterstrichen werden, dass in der Studie A nur sehr kurze Skalen verwendet werden konnten. Die in Studie A eingesetzten Skalen können dadurch nicht in gleicherweise die komplexe Struktur des AVEM abbilden und inhaltliche Verschiebungen sind folglich zu erwarten.

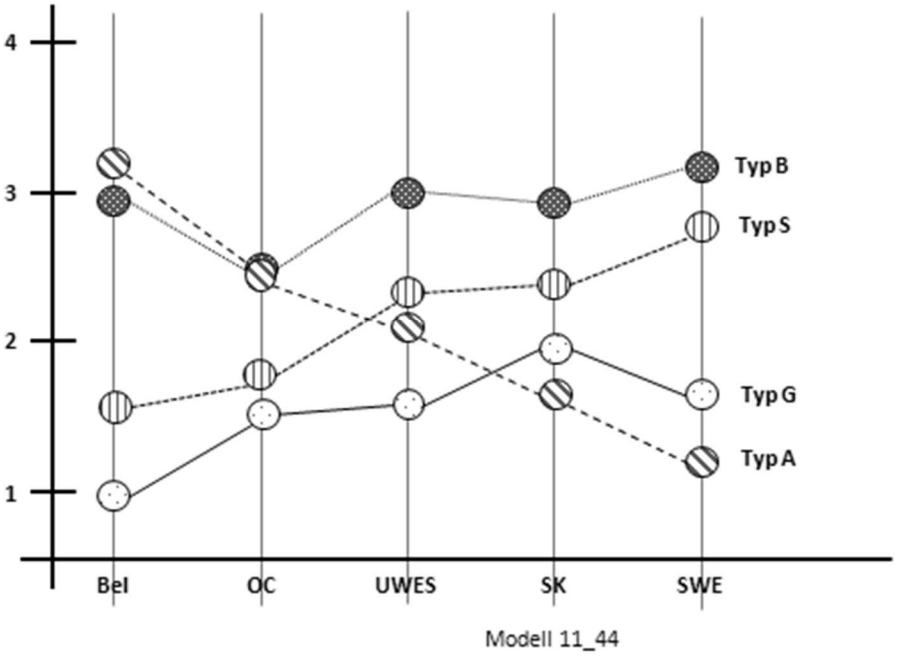


Abbildung 7.10: inhaltliche Darstellung der Klassen zum Modell M11 für vier Klassen – Klassenmittelwerte abgetragen

Im Vergleich der Skalen des AVEM im Bereich Arbeitsengagement mit den Items der UWES zeigt sich, dass in letzterem die Anstrengungsbereitschaft mit positiven Emotionen verknüpft ist, während im AVEM der Fokus auf einer Verausgabung liegt. Man könnte daher annehmen, dass Lehrkräfte, die in der vorliegenden Studie als Typ A klassifiziert wurden, bei ihrem Engagement diese positiven Emotionen nicht erleben und die Verschiebung zum originalen Muster daher entstammt. Aus dieser Überlegung heraus wurden auch Klassenlösungen für das Modell M11a berechnet, die als sechste Skala die Gewissenhaftigkeit berücksichtigt, deren Items stärker den verpflichtenden Gedanken der Arbeitsansprüche abdecken, wie sie auch im AVEM abgebildet werden.

Die Kennzahlen zu diesem Modell zeigen erfreulicherweise deutlich akzeptablere Werte für die Vier-Klassenlösung und die Fünf-Klassen-Lösung und liegen jeweils innerhalb der Intervalle zu 90% (LQT und CHI). Nur für die Drei-Klassen-Lösung weichen die Werte für den LQT und den CHI wiederum deutlich von den simulierten Bootstrape-Verteilungen ab. Erneut deutet der AIC-Wert auf eine bessere Passung der Fünf-Klassenlösung hin, wohingegen der BIC-Wert für eine Vier-Klassenlösung spricht. Die durchschnittliche Zuordnungswahrscheinlichkeit ist für die Vier-Klassenlösung am größten. Die Entscheidung, wiederum die Vier-Klassenlösung für die Interpretation und auch für die weitere Analyse zu wählen, ist auch inhaltlich getroffen worden, weil somit zumindest im Ansatz die Analogie zur Klassifikation des AVEM aufrechterhalten werden kann.

Tabelle 7.19

Kennzahlen zum Modell M11a mit vierstufigen Skalen-Items

Klassenzahl	Parameter	L	AIC	BIC	LQT	CHI	Zuordnungs-wahrschein-lichkeit
3	56	-2716.46	5544.923	5764.381	1283.911	4366.450	85.7%
4	75	-2683.02	5516.043	5809.960	1216.754**	3933.754**	86.8%
5	94	-2656.68	5501.351	5869.727	1164.410**	3553.827**	86.5%

Anmerkung: *:bei Niveau α=.05 nicht signifikant **: bei Niveau α=.10 nicht signifikant

Die in Abb. 7.5 dargestellte Vier-Klassenlösung des Modells M11a zeigt für die fünf aus dem Modell M11 übernommenen Skalen keine auffällige Abweichung. Entsprechend der Erwartung zeichnen sich Lehrkräfte, die ähnlich zur Klasse Typ A aus dem AVEM schienen, durch die größte Gewissenhaftigkeit aus. Lehrkräfte dieser Klasse legen wesentlich mehr Gewissenhaftigkeit an den Tag als Lehrkräfte, die dem Typ B‘ oder dem Typ S‘ zugeordnet

wurden, und auch ein wenig mehr als Lehrkräfte aus der Klasse Typ G‘. Tatsächlich staucht die Darstellung hier allerdings die Abstände, da sie auf Mittelwertbildung zugreift. Betrachtet man statistisch sinnvoller die möglichen Expertengrade für die Skala GW, erreichen hier 30.9% der Lehrkräfte aus der Klasse Typ A‘ die oberste Stufe, in den Klassen Typ B‘ und Typ S‘ sind es genau 16.0% bzw. 19.4%. Bei der Skala UWES erreichen immerhin 63.2% aus der Klasse Typ A‘ den höchsten Expertengrad, während es bei der Klasse Typ B‘ nur 21.3% sind. Die Lehrkräfte aus den beiden anderen Klassen kommen auf 86.5% (Typ G‘) und 57.1% (Typ S‘). Über beide Skalen zusammen genommen zeigen Lehrkräfte des Typs A‘ im Bereich des Arbeitsengagements die größten Werten.

Lehrkräfte, die als Typ A‘ klassifiziert wurden, haben einen Anteil an den klassifizierten Lehrkräften von 18.3% und zeichnen sich folglich durch ein erhöhtes Arbeitsengagement und durch eine deutliche Diskrepanz zwischen ihren auf fachdidaktische Herausforderungen bezogenen starken Kompetenz- und Kontrollüberzeugungen einerseits und die mit ihrer Tätigkeit verbundenen beruflichen Emotionen (erlebte Beanspruchung und Distanzierungsfähigkeit). Lehrkräfte dieses Typs erleben sich möglicherweise situativ als sehr wirkungsvoll, sehen sich aber insgesamt Ansprüchen gegenüber gestellt, denen gegenüber sie sich doch nicht gewachsen fühlen. Wer zu diesem Muster gehört, ist gefährdet, durch Engagement über seinen Möglichkeiten „auszubrennen“, es existiert für diese Lehrkräfte ein Defizit zwischen eingebrachtem Engagement und ausgleichendem Wohlbefinden.

Nicht ganz erwartungskonform ist das Muster des Typs G‘ mit einer Auftrittswahrscheinlichkeit von 19.9%. Einerseits sind diesem Muster nur Lehrkräfte zugeordnet worden, die den höchsten Expertengrad bei der erlebten beruflichen Beanspruchung aufweisen, andererseits zeigen sie nur die zweithöchsten Werte bei den Kompetenz- und Kontrollüberzeugungen. Warum sich die Lehrkräfte dieses Typs nur relativ geringe Kompetenzen und Kontrolle in fachdidaktischen Fragen zuschreiben, kann hier nicht aufgeklärt werden. Es ist aber zu berücksichtigen, dass der Abstand zwischen der höchsten und der zweithöchsten Expertenstufe bei einigen Items der Skalen SK und SWE nur minimal ist und dadurch der tatsächliche Unterschied zwischen den Muster Typ G‘ und Typ A‘ in Abb. 7.11 verzerrt ist.

Auf das Muster Typ B‘ entfallen 20.2% der klassifizierten Lehrkräfte. Auffällig ist bei diesem Muster der bereits beschriebene relativ niedrige Wert bei der Gewissenhaftigkeit als Variante des Arbeitsengagements. In dieser Klasse befinden sich daher Lehrkräfte, die mit ihren Aufgaben wohlmöglich tatsächlich momentan überfordert sind. Aber auch ausgebrannte Lehrkräfte, die sich in der Schule einbringen möchten, aber weder in Bezug auf die Schüler noch auf ihr eigenes Wohlbefinden Erfolgserlebnisse erhalten, sind dieser Klasse zugeordnet. Im Vergleich zur originalen Risikogruppe Typ B des AVEM zeigen sich auch mit Hinzunahme der Skala GW deutliche Parallelen zum originalen Muster.

Als Typ S‘ wurden 41.7% der Lehrkräfte klassifiziert. Erwartungskonform zeigen Lehrkräfte aus dieser Klasse eine geringe Neigung, sich gewissenhaft einzubringen. Die von rechts nach links gesehen fallende Kurve verdeutlicht, dass sie dieses geringe Engagement aber in der

Tat zumindest für sich selbst sinnvoll nutzen können. Wie für die als Typ G‘ klassifizierten Lehrkräfte sollten für Lehrkräfte des Typs S‘ die größte Passung zu den in Kap. 5 formulierten Hypothesen zu erwarten sein. Das Modell M11a eignet sich mit der Vier-Klassenlösung allerdings insgesamt relativ gut für weitere Analyse, inwieweit sich das Vorbereitungsverhalten als konditionalprogrammierendes Vorgehen auffassen lässt, und wird in Abs. 7.3 daher auch für weitere Analysen verwendet.

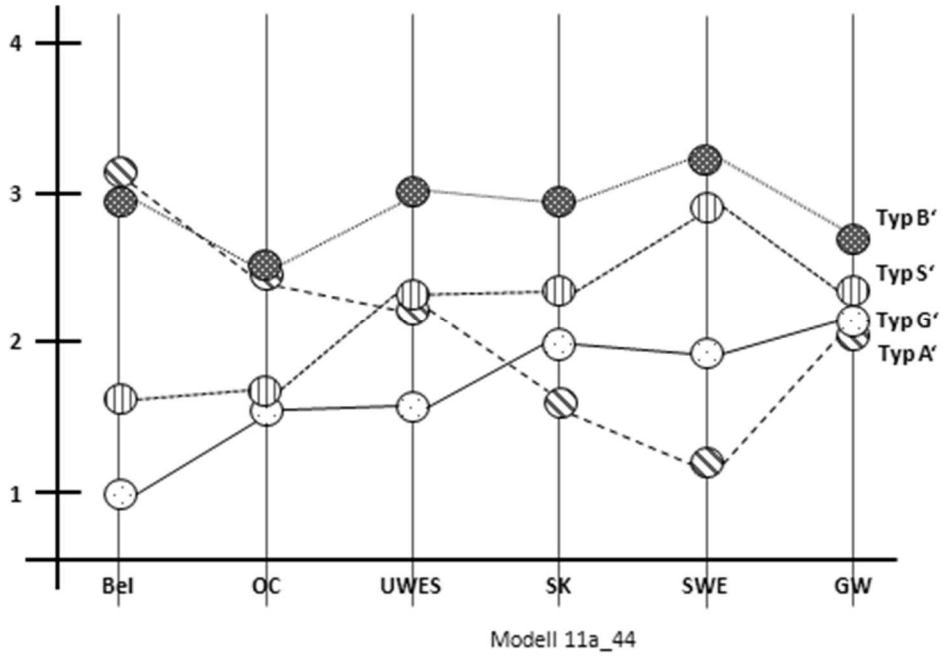


Abbildung 7.11: inhaltliche Darstellung der Klassen zum Modell M11a für vier Klassen – Klassenmittelwerte abgetragen

Abb. 7.12 zeigt die Fünf-Klassenlösung. Dabei bleiben die vier Klassen der Vier-Klassenlösung bestehen und werden um eine Klasse ergänzt (ungemusterte Kreise), die schwerlich zu interpretieren ist. Lehrkräfte, die dieser Klasse zugeordnet wurden, zeigen die geringsten Werte auf der Skala UWES, zeichnen sich bei den fünf anderen Skalen aber durch durchschnittliche Werte aus. Die Klasse besitzt im Verlauf der linken Hälfte eine gewisse Ähnlichkeit mit der Klasse Typ B‘, gleichzeitig besitzen Lehrkräfte dieser Klasse offensichtlich eine etwas positivere Kontrollüberzeugung als Lehrkräfte des Typs B‘ und zeigen wesentlich mehr Gewissenhaftigkeit ähnlich zu dem Typ A‘.

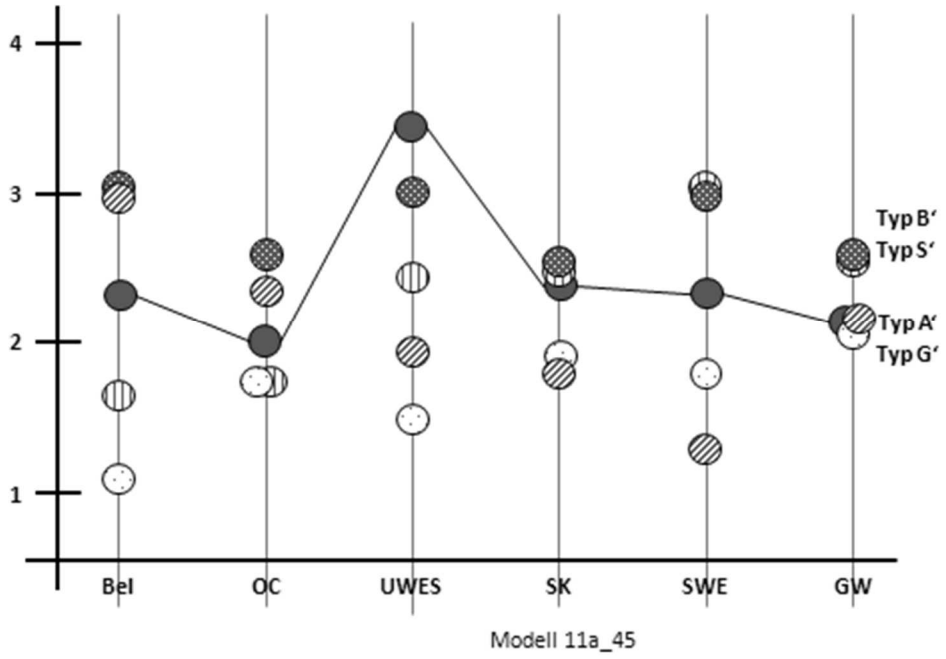


Abbildung 7.12: inhaltliche Darstellung der Klassen zum Modell M11a für fünf Klassen – Klassenmittelwerte abgetragen

Modelle mit personenbezogene Überzeugungen und Ziel- und Kausalüberzeugungen über das Lehren & Lernen

Als zweite Modellgruppe wurden latenten Klassenanalysen durchgeführt, bei denen neben den Skalen der Modell M11 und M11a auch Überzeugungen über das Lehren & Lernen berücksichtigt wurden. Die dabei integrierten Skalen waren die sechs Items umfassende Skala Konstruktionsüberzeugung (KÜ) und die fünf Items umfassende Skala Transmitterüberzeugung (TÜ). Für das Modell M12 wurde das Modell 11 eben genau um diese beiden Skalen erweitert. Die Skalen KÜ und TÜ wurden als fünfstufige Skalen integriert, obwohl der Vergleich der AIC-Werte eine zweistufige (KÜ) bzw. dreistufige (TÜ) Einteilung nahelegt. Für das Modell M12 wurden wiederum Drei-, Vier- und Fünf-Klassen-Lösungen berechnet. Tab 7.20 Kann man entnehmen, dass das Drei-Klassenlösung des Modells nicht auf die Daten passt. Für die beiden anderen Varianten sind die Kennzahlen für den CHI im akzeptablen Bereich, aber nur die Vier-Klassenlösungen weist für den LQT zumindest einen akzeptablen Wert auf.

Tabelle 7.20

Kennzahlen zum Modell M12 mit vierstufigen Skalen-Items

Klassenzahl	Parameter	L	AIC	BIC	LQT	CHI	Zuordnungs-wahrschein-lichkeit
3	65	-3013.50	6157.003	6411.906	1785.927	23587.62	85.6%
4	87	-2979.46	6132.921	6474.099	1713.136*	15443.35**	87.5%
5	109	-2948.90	6115.791	6543.243	1655.115	12331.66**	87.5%

Anmerkung: *:bei Niveau α=.05 nicht signifikant **: bei Niveau α=.10 nicht signifikant

Abb. 7.13 zeigt die Struktur dieser Vier-Klassenlösung. Dabei fällt als erstes auf, wie eng die vier Klassen bei den Skalen KÜ und TÜ zusammenliegen. Die beiden Skalen haben offensichtlich wenig zur Zuweisung der Lehrkräfte in die beiden Klassen beigetragen. Die Skala KÜ zeigt sich mit einer weiteren Spannweite zwischen den Klassen als die Skala TÜ, allerdings wurde der Skala KÜ eine vierstufige Struktur auferlegt, obwohl die Daten maximal eine zweistufige Struktur nahelegen. Die Interklassenhomogenität bzgl. dieser Skala ist folglich in Wirklichkeit viel größer als sie hier erkennbar ist. Unter diesem Blickwinkel sind auch die inkonsistenten Anordnungen der ersten, dritten und vierten Klasse bei diesen beiden Skalen zu sehen, denn obwohl die beiden Skalen im Prinzip das gleiche Konstrukt messen, ändert sich die Reihenfolge der Anordnung stark.

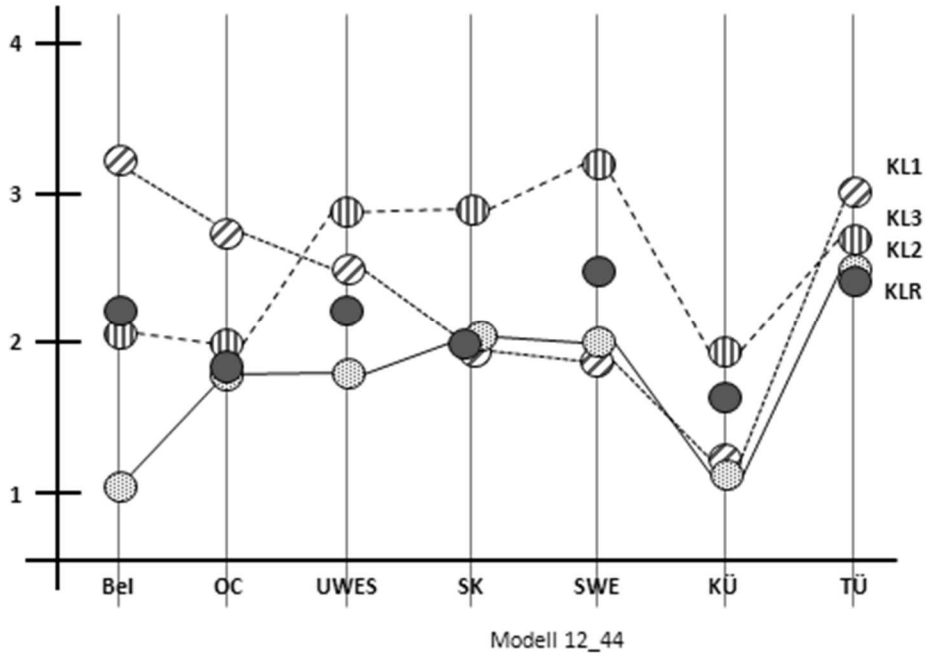


Abbildung 7.13: inhaltliche Darstellung der Klassen zum Modell M12 für vier Klassen – Klassenmittelwerte abgetragen

Über alle sieben Skalen betrachtet zeigt die erste Klasse (KL 1) einen Verlauf, der der Klasse Typ A aus dem Modell M11 ähnlich ist. Lehrkräfte dieses Typs zeichnen sich durch eine relativ positive Kontroll- und Kompetenzüberzeugung aus, zeigen aber beim Arbeitsengagement bzw. der Arbeitszufriedenheit schlechtere Werte und grenzen sich insbesondere bei den Skalen OC und BEL negativ von den anderen Klassen ab. Demgegenüber weist die zweite Klasse (KL 2) eine Parallele zur Klasse Typ G auf. Lehrkräfte haben ähnliche Werte bei den Skalen SK und SWE, aber auch jeweils die besten Werte in den drei linken Skalen. In der dritten Klasse sind diejenigen Lehrkräfte gruppiert, die sich konträr zur ersten Klasse zeigen und der Klasse Typ S gleichen. Relativ gute Werte bei der erlebten Belastung und dem Overcommitment stehen negative Kompetenz- und Kontrollüberzeugungen gegenüber, auch der Wert bei der Skala UWES ist sehr negativ. Während Lehrkräfte aus der ersten und zweiten Klasse eindeutig mehrheitlich die höchstmögliche Zustimmung zu einem konstruktivistischen Verstehen von Mathematik zeigen, ist für Lehrkräfte der dritten Klasse (Kl 3) eine weniger deutliche Tendenz zu erkennen. Noch einmal sei aber darauf hingewiesen, dass die vierstufige Struktur bei dieser Skala künstlich ist und minimale Unterschiede überbetont. Die vierte Klasse (KL R) zeichnet sich in den Skalen durch mittlere Werte aus. Nur bei der Skala TÜ haben Lehrkräfte dieser Klasse die positivsten Werte, unterscheiden sich aber kaum merklich von den Werten der zweiten und dritten Klasse. Damit fehlt in dieser Klasseneinteilung ein Äquivalent zur Klasse Typ B. Auch ist die Verteilung auf die vier Klassen wesentlich anders als bei dem Modell M11.

In der ersten Klasse befinden sich 22.0% der 372 klassifizierten Lehrkräfte, die zweite Klasse umfasst 29.2%, die dritte 35,4% und die vierte Klasse nur 13.1%. Insofern handelt es sich bei der vierten Klasse möglicherweise um eine Gruppe von unskalierbaren und damit um eine inhaltliche Dreiklassenlösung. In jedem Fall liegt keine tatsächliche Nähe zum Modell M11 vor.

Da die Skala KÜ nur schwach differenziert, scheint es nicht lohnenswert, diese in weitere Analysen einzubeziehen. Mit den Überlegungen zu Modell M11a und der etwas besseren Streuung der Skala TÜ ist eine Erweiterung des Modells M11a um die Skala TÜ denkbar. Das Modell M12a besteht dann aus den Skalen BEL, OC, UWES, SK, SWE, GW und TÜ. Für die Transmitterüberzeugungen wird allerdings dieses Mal eine dreistufige Einteilung berücksichtigt, weil diese Einteilung der Datenstruktur besser gerecht wird.

Die Kennzahlen für dieses Modell ergeben (nur knapp) akzeptable Werte für den CHI und LQT für die Vier- und Fünf-Klassen-Lösung. Die Kennwerte sprechen dabei eher für die Fünf-Klassen-Lösung. Wie Tab. 7.21 darüber hinaus entnommen werden kann, legt der AIC-Vergleich die Fünf-Klassen-Lösung, der BIC-Vergleich wiederum die Vier-Klassen-Lösung nahe.

Tabelle 7.21

Kennzahlen zum Modell M12a mit vierstufigen Skalen-Items

Klassenzahl	Parameter	L	AIC	BIC	LQT	CHI	Zuordnungs-wahrschein-lichkeit
3	62	-3078.02	6280.046	6523.018	1871.768	13479.06	85.2%
4	83	-3044.85	6255.693	6580.619	1805.067*	12811.99*	86.7%
5	104	-3014.99	6237.973	6645.538	1748.939*	11281.08**	85.7%

Anmerkung: *:bei Niveau α=.05 nicht signifikant **: bei Niveau α=.10 nicht signifikant

Vergleicht man die Vier-Klassen-Lösung (Abb. 7.14) und die Fünf-Klassen-Lösung (Abb. 7.15) inhaltlich, kann man in der Vier-Klassen-Lösung eine große Übereinstimmung mit dem Modell M11a erkennen und auch in der Fünf-Klassen-Lösung lassen sich die vier Klassen identifizieren, die auch im Modell M11a abgebildet werden. Zusätzlich beinhaltet die Fünf-Klassen-Lösung eine kleine Gruppe von dreißig Lehrkräften, die bei den Skalen OC, SK, SWE und GW sehr ähnliche Überzeugungen besitzen wie die Lehrkräfte aus der Klasse Typ S‘, aber bei der beruflich erlebten Beanspruchung und beim Arbeitsengagement bzw. der

Arbeitszufriedenheit negativere bzw. sogar wesentlich negativere Werte aufweisen. Auch zeichnet sich diese Gruppe durch sichtbar größere Transmitterüberzeugungen aus, während alle anderen Gruppen hier ähnlich positive Werte zeigen. Für die ursprünglich im M11a unterschiedenen Klassen fungiert die Skala TÜ folglich nicht als zusätzliches Differenzierungskriterium. Die personenbezogenen Überzeugungen zeigen sich hier von Überzeugungen über das Lehren & Lernen von Mathematik unabhängig. Auf eine detaillierte Beschreibung der beiden Klassenlösungen zu Modell M12a wird daher an dieser Stelle verzichtet, da dieses Modell weder vorher aus vorherigen Befunden hergeleitet werden konnte noch die angepasst Skala TÜ sich als sinnvolles Differenzierungskriterium zeigt.

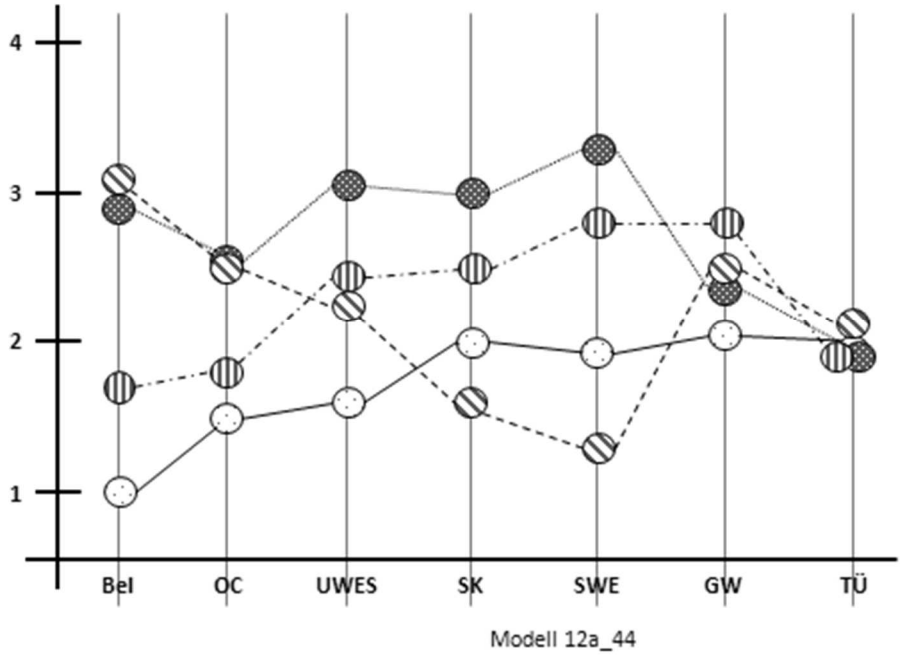


Abbildung 7.14: inhaltliche Darstellung der Klassen zum Modell M12a für vier Klassen – Klassenmittelwerte abgetragen

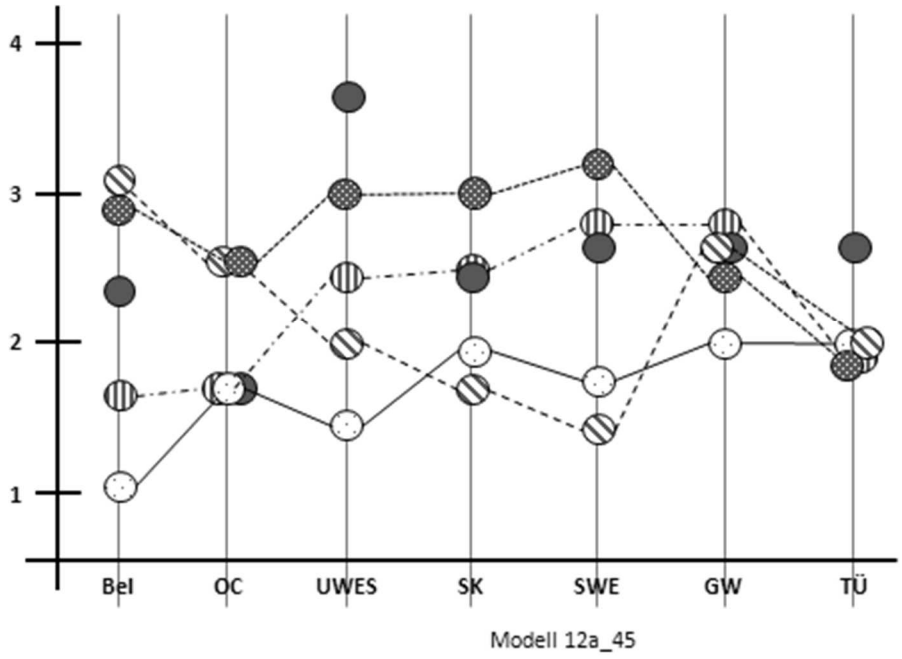


Abbildung 7.15: inhaltliche Darstellung der Klassen zum Modell M12a für fünf Klassen – Klassenmittelwerte abgetragen

Damit kann für den vorliegenden Datensatz einzig das Modell M11a aus der ersten Kategorie als für weitergehende Analysen des Vorbereitungsverhaltens sinnvolles Modell gelten. Im nächsten Abschnitt werden nun die Modelle diskutiert, bei denen vorausgesetzt wird, dass zentrale Vergleichsarbeiten von den befragten Lehrkräften als ein irgendwie geartetes Feedbackinstrument gesehen werden.

7.2.2 Modelle auf Grundlage der Feedback Intervention-Theorie von Kluger und DeNisi sowie nachfolgender Forschungsbefunde als Prädiktoren des Vorbereitungsverhaltens

Ein Modell auf Grundlage der Feedback Intervention-Theorie von Kluger und DeNisi ohne direkten VERA8-Bezug als Prädiktoren des Vorbereitungsverhaltens

Wie in Kap. 6 erläutert wurde, gehören der Umfang an eigenen Ressourcen wie die Selbstwirksamkeitserwartung (SWE) und die Gewissenhaftigkeit (GW), die Attribution der Ergebnisse (LA) und Internalisierung der Ziele (hier für die Teilkompetenzen „Argumentieren & Kommunizieren“ [ZieleArK], „Problemlösen & Modellieren“ [ZielePLM] und „Werkzeuge nutzen“ [ZieleW]) zu den Variablen, die dem theoretischen Rahmen folgend den Umgang mit

Feedback beeinflussen. Diese Bedingungen sind im Modell M21 zusammengefasst. Weiter wurde erläutert, dass vier verschiedene Muster zu erwarten sind. Diese ergeben sich daraus, ob sowohl die Ziele akzeptiert und die Testergebnisse auf die eigene Leistung zurückgeführt werden als auch ausreichend eigene (oder externe) Ressourcen zur Verfügung stehen, um den Feedbackprozess erfolgreich zu nutzen.

Es wurden bei den drei Ziel-Skalen dreistufige, ansonsten vierstufige Skalen verwendet. Wie bereits vorher auch wurden Drei-, Vier- und Fünf-Klassen-Lösungen berechnet, deren Kennzahlen in Tab. 7.22 angegeben sind. Insgesamt wurden n=371 Lehrkräfte klassifiziert. Auffällig ist, dass alle drei Klassenlösungen offensichtlich zum Datensatz passen. Auch die Zuordnungswahrscheinlichkeit ist für alle drei Klassenlösungen ähnlich hoch. Der Vergleich des AIC lässt anders als theoretisch hergeleitet eine Präferenz für die Fünf-Klassenlösung erkennen. Nachfolgend werden daher die Vier- und die Fünf-Klassen-Lösung inhaltlich diskutiert.

Tabelle 7.22

Kennzahlen zum Modell M21 mit drei- bzw. vierstufigen Skalen-Items

Klassenzahl	Parameter	L	AIC	BIC	LQT	CHI	Zuordnungs-wahrschein-lichkeit
3	47	-2380.21	4874.362	5038.621	776.857**	1631.487**	88.4%
4	63	-2361.81	4849.619	5096.509	739.3161**	1468.175**	89.8%
5	79	-2343.67	4845.33	5154.922	705.0385**	1400.283**	87.4%

Anmerkung: *:bei Niveau α=.05 nicht signifikant **: bei Niveau α=.10 nicht signifikant

Es zeigt sich, dass das Vier-Klassen-Modell keine sich ergänzenden Muster besitzt, sondern vorwiegend Niveaustufen ausweist. Dabei sind die zweite, dritte und vierte Klasse ([KL2], [KL3] und [KL4]) geordnet, während die erste Klasse (KL1) bei der Skala LA nur den drittbesten Wert aufweist und auch die Ordnung bei der Skalen zur Selbstwirksamkeits-erwartung (KL1 M=2.20, KL2 M=2.19) verletzt. Bei den drei Ziel-Skalen bleibt die Ordnung hingegen erhalten, wenn auch die Ziele innerhalb der vier Klassen unterschiedlich gewichtet sind. KL1 und KL2 umfassen jeweils Lehrkräfte, die vor allem das Argumentieren & Kommunizieren als besonders wichtiges Ziel ihres Unterrichts angeben und die Nutzung von mathematischen Werkzeugen weniger Gewicht einräumen. Lehrkräfte der dritten Klasse scheinen alle drei Kompetenzen hingegen durchschnittlich zu gewichten und Lehrkräfte der vierten Klasse bevorzugen vor allem das Problemlösen & Modellieren. In der ersten Klasse

sind mit n=174 (46.9%) die meisten Lehrkräfte zusammengefasst, die vierte Klasse umfasst mit n=16 (4.3%) nicht einmal ein Zehntel dieser Größe. Die zweite Klasse gruppiert n=64 (17.3%) und die dritte n=117 (31.5%) Lehrkräfte.

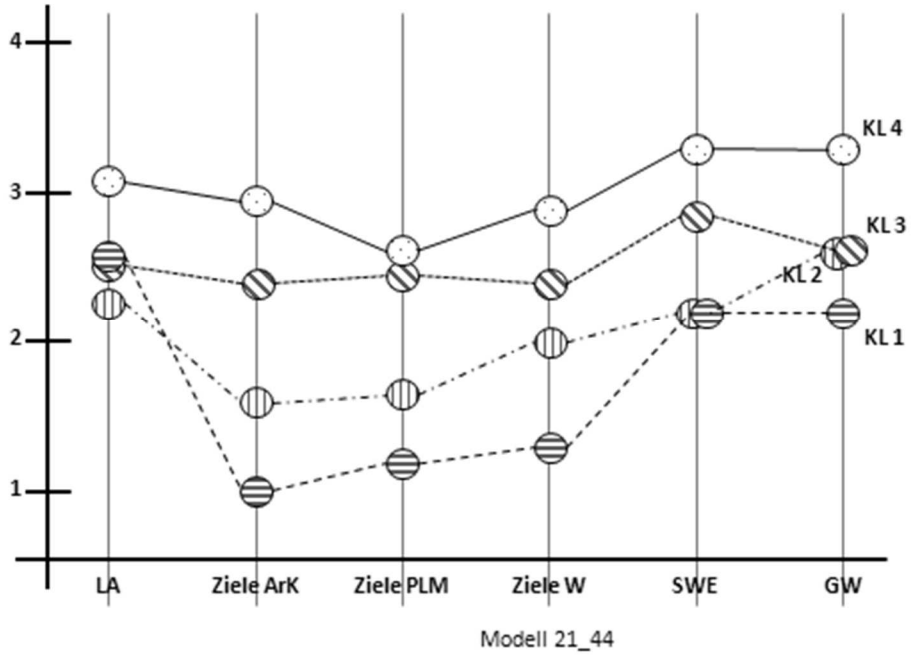


Abbildung 7.16: inhaltliche Darstellung der Klassen zum Modell M21 für vier Klassen – Klassenmittelwerte abgetragen

In der grundsätzlichen Tendenz bleiben die vier Klassen aus der Vier-Klassen-Lösung auch in der Fünf-Klassen-Lösung erhalten. Lediglich bei der Skala GW kreuzen sich die Muster der zweiten und dritten Klasse, sodass auch hier die Ordnung verletzt wird. Die Verteilung auf die vier Klassen ändert sich durch die fünfte Klasse besonders für die beiden größten Klassen. Die erste Klasse umfasst nur noch n=148 (39.9%) Lehrkräfte und die dritte nur noch n=104 (28.0%), in der zweiten Klasse sind nun n=70 (18.9%) Lehrkräfte gruppiert und die vierte Klasse ist mit n=18 (4.9%) auch minimal größer geworden.

Die fünfte Klasse mit n=31 (8.4%) besitzt ein davon stark abweichendes Muster. Während Lehrkräfte dieses Typs die Leistung von Schülern eher nicht als Resultat ihrer eigenen Arbeit betrachten, also eher negative Werte auf der Skala LA zeigen, und die schlechtesten Werte bei der Skala GW aufweisen, besitzen sie durchaus eine durchschnittliche Selbstwirksamkeitserwartung und geben an, die erhobenen Ziele auch als relevant für ihren Unterricht zu halten.

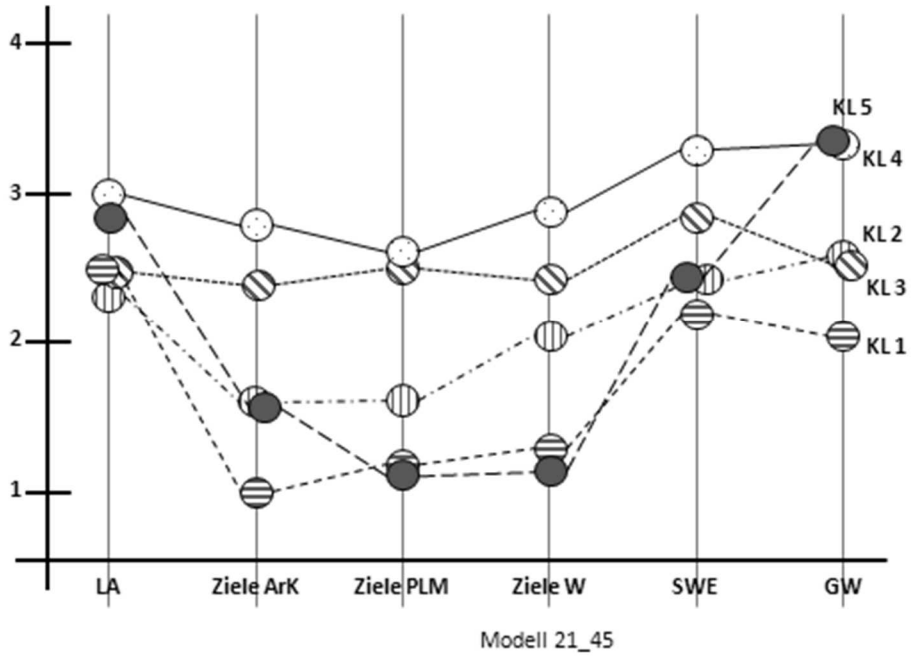


Abbildung 7.17: inhaltliche Darstellung der Klassen zum Modell M21 für fünf Klassen – Klassenmittelwerte abgetragen

Die Interpretation des fünften Musters im Zusammenhang mit den vier anderen Mustern ist dadurch nicht in der für weitere Analysen nötigen Einfachheit möglich. Es fehlt in beiden Klassenlösung an einem Muster, welches insgesamt gute Bedingungen für die Nutzung von Feedback ausweist. Jenes Muster ist als Referenzrahmen allerdings notwendig. Die Vier- und Fünf-Klassen-Lösung eignen sich folglich eher weniger für eine Analyse des Vorbereitungsverhaltens.

Ein Modell auf Grundlage der Feedback Intervention-Theorie von Kluger und DeNisi unter Berücksichtigung des Evaluationszyklus‘ nach Helmke als Prädiktoren des Vorbereitungsverhaltens

Das fünfte und letzte Modell (M22) basiert ebenfalls auf der Feedback Intervention-Theorie von Kluger und DeNisi, berücksichtigt aber explizit den Umgang mit VERA8-Ergebnisse durch die Lehrkräfte in den vergangenen Jahren. Als einziges Modell basiert es auf dem Datensatz der Studie B und die Schätzungen der Klassenlösungen können sinnvollerweise auch nur diejenigen Lehrkräfte einbeziehen, die bereits vorher Erfahrungen mit VERA8 gesammelt haben, denn nur diese können über ihren vorherigen Umgang mit VERA8-Ergebnissen berichten.

Konkret umfasst das Modell M22 im Unterschied zu Modell M21 Skalen zur Rezeption von vorherigen VERA8-Ergebnissen (REPZ), deren Reflexion (RFL) und das Bestreben, daraus ggf. Unterrichtsveränderungen abzuleiten (VEA). Diese Skalen treten in Summe auch an die Stelle der Skala LA. Die Attribution der Testleistung wird folglich nur indirekt aus dem Umgang mit den VERA8-Ergebnissen geschlossen. Wiederum werden die Gewissenhaftigkeit (GW), die fachdidaktische Selbstwirksamkeitserwartung (SWE) und die Akzeptanz der vorgegebenen Unterrichtsziele in das Modell integriert. Die Akzeptanz der vorgegebenen Unterrichtsziele wurde aber für die Studie B nur über eine allgemeine Akzeptanz der Kernlehrpläne (KL) erfasst. Schließlich fand auch eine Skala zur erlebten Unterstützung durch die Schulleitung (USL) Eingang in das Modell, sodass zusammen mit der Gewissenhaftigkeit und Selbstwirksamkeitserwartung beide potenzielle Ressourcenarten für eine erfolgreiche Nutzung von Feedback berücksichtigt worden sind. Alle sieben Skalen sind in diesem Fall vierstufig.

Die Kennzahlen zu Modell M22 sind Tab. 7.23 zu entnehmen. Insgesamt konnten n=198 Lehrkräfte mittels dieses Modells klassifiziert werden. Für die Vier- und die Fünf-Klassenlösung ergibt sich jeweils eine ausgezeichnete Passung zum Datensatz, für die Drei-Klassenlösung gilt dies nicht. AIC und BIC zeigen beide an, eher die Vier-Klassenlösung statt der Fünf-Klassenlösung zu präferieren. Die Zuordnungswahrscheinlichkeit ist allerdings für die Fünf-Klassenlösung (92.4%) minimal größer als für die Vier-Klassenlösung (92.0%).

In Kap. 5 ist hergeleitet worden, dass vier Typen bei der Nutzung von Feedback zu erwarten sind. Mit den Überlegungen zu den Ergebnissen der Studie von Hosenfeld (Hosenfeld, 2010) splitten sich diejenigen, die alle Voraussetzungen besitzen, um das durch VERA8 angebotene Feedback zu nutzen, noch einmal in zwei Untertypen auf: je nachdem ob sie einen Veränderungsbedarf feststellen oder ein positives Feedback erhalten. Auch hier werden erst einmal beide Klasseneinteilungen, die Vier- und die Fünf-Klassenlösung, inhaltlich vorgestellt. Einen Überblick geben wiederum die beiden Abb. 7.18 und 7.19. Zusätzliche Hinweise geben aber in diesem Fall stellenweise auch die Modalwerte, die in der Graphik nicht dargestellt werden können.

Die erste Klasse der Vier-Klassenlösung besteht aus n=50 Lehrkräften (25.2%) und kann in Anlehnung an die Klassifikation von Stamm als Typ „Blockierer“[128] (Typ b[129]) bezeichnet werden. Lehrkräfte dieses Typs zeichnen sich dadurch aus, dass sie mittlere bis wenig eigene Ressourcen (GW und SWE) besitzen, aber den neuen Kernlehrplänen vor allem mit wenig

[128] Die Bezeichnung als „Blockierer“ resultiert aus dem berichteten Umgang mit den Ergebnissen aus vorherigen VERA8-Durchgängen und dient der einfacheren Beschreibung. Genau genommen kann nicht belegt werden, dass die derart klassifizierten Lehrkräfte tatsächlich aktiv die Evaluation mit VERA8 blockieren. Sie geben lediglich an, die ihnen zur Verfügung gestellten Daten nicht zu nutzen und an ihrer Nutzung nicht interessiert zu sein. Durch den Bezeichnung „Blockierer“ implizit ausgedrückte Kausalzusammenhänge zwischen den einzelnen Skalen werden an dieser Stelle weder inhaltlich noch methodisch unterstellt. Damit sind weder Folgerungen der Art, „diese Lehrkräfte nutzen die VERA8-Daten bewusst nicht, weil sie sich gegen die Kernlehrpläne wehren“, noch der Art „diese Lehrkräfte haben zu wenig Ressourcen, um an einem Feedbackprozess teilzunehmen, und lehnen VERA8 *deswegen* ab“, zulässig.

[129] Zur besseren Unterscheidung der Klassen vom Modell M11a werden Kleinbuchstaben genutzt.

Akzeptanz begegnen. Abweichend von der graphischen Darstellung nehmen einige dieser Lehrkräfte außerdem am wenigsten eine Unterstützung durch die Schulleitung wahr, während andere Lehrkräfte aus dieser Gruppe dies genau anders erleben. Der Modalwert für die Skala USL beträgt X_d=1, während der mittlere Wert bei M=2,54 liegt. Gleichzeitig offenbaren sie die negativsten Werte für die drei VERA8-Skalen zu Rezeption, Reflexion und möglichen Veränderungen. Nach der Klassifikation aus Kap. 5 ist diese Gruppe als F21.C einzustufen.

Tabelle 7.23

Kennzahlen zum Modell M22 mit vierstufigen Skalen-Items

Klassenzahl	Parameter	L	AIC	BIC	LQT	CHI	Zuordnungs-wahrschein-lichkeit
3	65	-1636.85	3403.702	3617.439	1182.979**	18711.65	90.9%
4	87	-1607.68	3389.335	3675.434	1127.179**	13331.67**	92.0%
5	109	-1586.593	3391.187	3749.608	1088.965**	11848.50**	92.4%

Anmerkung: *:bei Niveau α=.05 nicht signifikant **: bei Niveau α=.10 nicht signifikant

Die zweite Klasse besteht aus Lehrkräften, die ebenfalls wenig Akzeptanz für die neuen Kernlehrpläne zeigen und mittlere Werte im Bereich der eigenen Ressourcen besitzen. Lehrkräfte dieses Typs erleben im Unterschied zu den in Klasse Typ b zusammengefassten Lehrkräften aber insgesamt auch eine Unterstützung der Schulleitung. In der theoretisch hergeleiteten Einteilung aus Kap. 5 würde die Klasse als F21.D eingestuft. Sie haben sich bisher nur bedingt mit den Ergebnissen aus VERA8 beschäftigt, geben aber an, trotzdem aus diesen Veränderungen ableiten zu wollen. Dieses auf den ersten Blick paradoxe Phänomen tritt damit parallel zu den Befunden von Hosenfeld auf. Insgesamt lässt sich dieses Muster wohl durchaus treffend als „Alibi“-Muster bezeichnen (Typ a), wobei die Bezeichnung aus der großen Differenz zwischen dem als Gewissenhaftigkeit bezeichnetem Engagement und der angegebenen Veränderungsbereitschaft auf der einen und der geringen Akzeptanz der Kernlehrpläne auf der anderen Seite resultiert. Dem Muster gehören n=58 (29.3%) Lehrkräfte an.

Lehrkräfte mit der größten Akzeptanz der Kernlehrpläne finden sich im Muster der dritten Klasse. Die nur n=22 Lehrkräfte (11.1%) bringen darüber hinaus große eigene Ressourcen mit (GW und SWE) und erleben auch eine große Unterstützung durch die Schulleitung. Sie haben sie am intensivsten mit den Ergebnissen auseinandergesetzt, folgern aus den Ergebnissen

aber mehrheitlich keinen Veränderungsbedarf. Der durchschnittliche Wert M=2.50 verzerrt die Diskrepanz innerhalb dieser Klasse bei der Skala VEA, denn der Modalwert liegt hier bei X_d=4. Das Muster umfasst somit Lehrkräfte, die durchaus die Ergebnisse aus VERA8 ihrem Zweck nach nutzen. Die Klasse wird daher als „Nutzung Eins“ (Typ nI) bezeichnet. Sie entspricht am meisten der theoretisch hergeleiteten Klasse F.21.A, d.h. dieser Klasse zugeordnete Lehrkräfte bringen die besten Voraussetzungen mit, um sich dem Feedbackprozess zu stellen. Gleichzeitig ist aber die Annahme nicht abwegig, dass sie mehrheitlich durch ihre persönlichen Ressourcen schon in der Lage sind, die durch die Kernlehrpläne bzw. Bildungsstandards gesetzten Ziele erfolgreich umzusetzen, sodass sie von dem Feedbackprozess selbst gar nicht mehr profitieren.

Einen ähnlichen Umgang mit den Ergebnissen aus vorherigen VERA8-Durchgängen zeigen auch Lehrkräfte der vierten Klasse. Dieses Muster wird folglich als „Nutzung Zwei“ (Typ nII) bezeichnet. Zwar ist die Rezeption und Reflexion nicht ganz so ausgeprägt wie bei Lehrkräften der dritten Klasse, die Werte sind aber trotzdem noch gut. Im Unterschied zu jenen Lehrkräften bringen Lehrkräfte des zweiten Nutzungstyps aber etwas weniger persönliche Ressourcen mit, insbesondere auf der Skala SWE sind die Werte nur mittelmäßig. Zum Teil erleben jene Lehrkräfte auch nur eingeschränkt eine ausreichende Unterstützung durch die Schulleitung. Dies gilt aber wiederum nicht für alle in dieser Klasse. Der Durchschnitt liegt zwar bei M=2.33, der Modalwert beträgt hier aber X_d=1. Im theoretisch hergeleiteten Modell aus Kap. 5 entspräche diese Klasse am ehesten der Klasse F.21.B. Auch ist die Akzeptanz der Kernlehrpläne hier auch nur durchschnittlich. Mit n=68 Lehrkräften (34.3%) handelt es sich bei dem vierten Muster um die größte Klasse. Zusammen mit einem Teil der Lehrkräfte aus der dritten Klasse sind Lehrkräfte des Typs nII diejenigen, die am wahrscheinlichsten von der Bereitstellung von VERA8-Daten profitieren: Sie besitzen auf der einen Seite ausreichend eigene Ressourcen und Akzeptanz für die Ziele, haben auf der anderen Seite womöglich aber auch noch genügend Verbesserungsbedarf, um sich dem Feedbackprozess zu stellen und auch neue Erkenntnisse aus den Daten zu gewinnen.

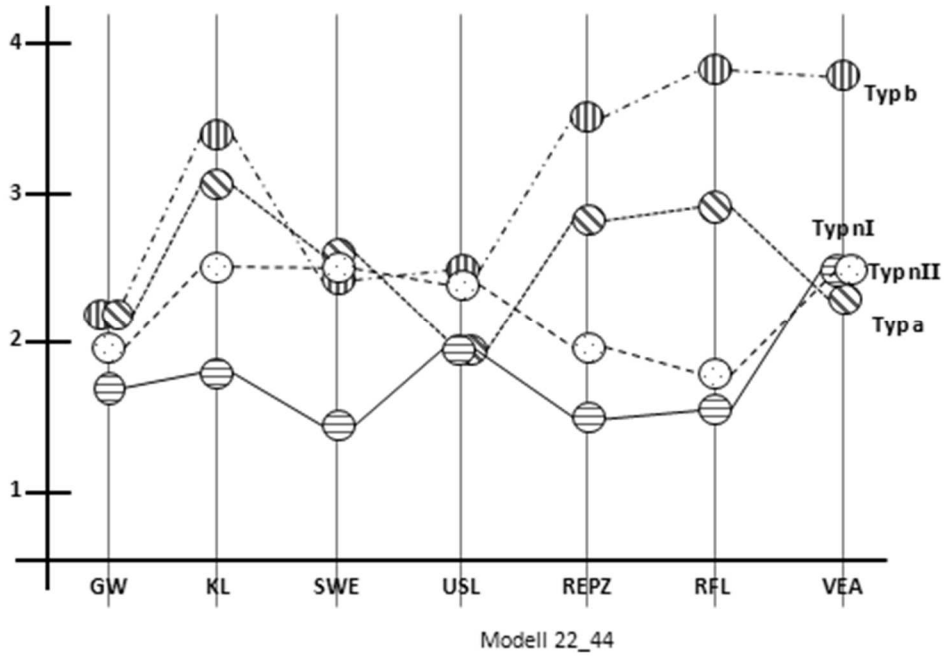


Abbildung 7.18: inhaltliche Darstellung der Klassen zum Modell M22 für vier Klassen – Klassenmittelwerte abgetragen

Im Vergleich zur Vier-Klassenlösung bleiben bei der Fünf-Klassenlösung die zweite, dritte und vierte Klasse als Muster erhalten. Es gibt hier lediglich leichte Verschiebungen, sodass sich die Größen der drei Klassen ändern. In der als „Typ a“ bezeichneten Klasse sind nun n=59 (29.8%) Lehrkräfte verortet, während in die als „Typ nI“ bezeichneten Klasse jetzt n=23 (11.6%) Lehrkräfte eingruppiert wurden. Die Klasse „Typ nII“ umfasst nur noch n=62 (31.3%) Lehrkräfte.

Die erste Klasse der Vier-Klassenlösung teilt sich hingegen in zwei etwa gleich große Klassen auf, die deswegen als „Blockade Eins“ (Typ bI) und „Blockade Zwei“ (Typ bII) bezeichnet werden. Zur Klasse „Typ bI“ gehören n=25 (12.6%) Lehrkräfte, die geringe eigene Ressourcen (GW, SWE) besitzen, die Kernlehrpläne nur wenig akzeptieren und nur eine minimale Auseinandersetzung mit VERA8-Ergebnissen angeben. Gleichzeitig berichten sie aber von einer großen wahrgenommenen Unterstützung durch die Schulleitung. Im Gegensatz dazu haben Lehrkräfte, die als „Typ bII“ gruppiert wurden, genügend eigene Ressourcen, stehen den Kernlehrplänen aber genauso kritisch gegenüber und nehmen die Schulleitung als wenig unterstützend wahr. Auch sie berichten nur über eine geringe Auseinandersetzung mit den VERA8-Ergebnissen. Zu dieser Klasse gehören n=29 (14.6%) der Lehrkräfte aus der Substichprobe. Für beide Gruppen kann folglich festgestellt werden, dass die dort klassifizierten Lehrkräfte nicht über ausreichend Ressourcen verfügen, um nach der FIT ausreichend für einen Feedbackprozess gerüstet zu sein. Der gleichzeitig hohe Grad der

Ablehnung der Kernlehrpläne bedeutet eine nicht ausreichende Akzeptanz derjenigen Zielvorgaben, deren Erreichen mit VERA8 evaluiert werden soll.

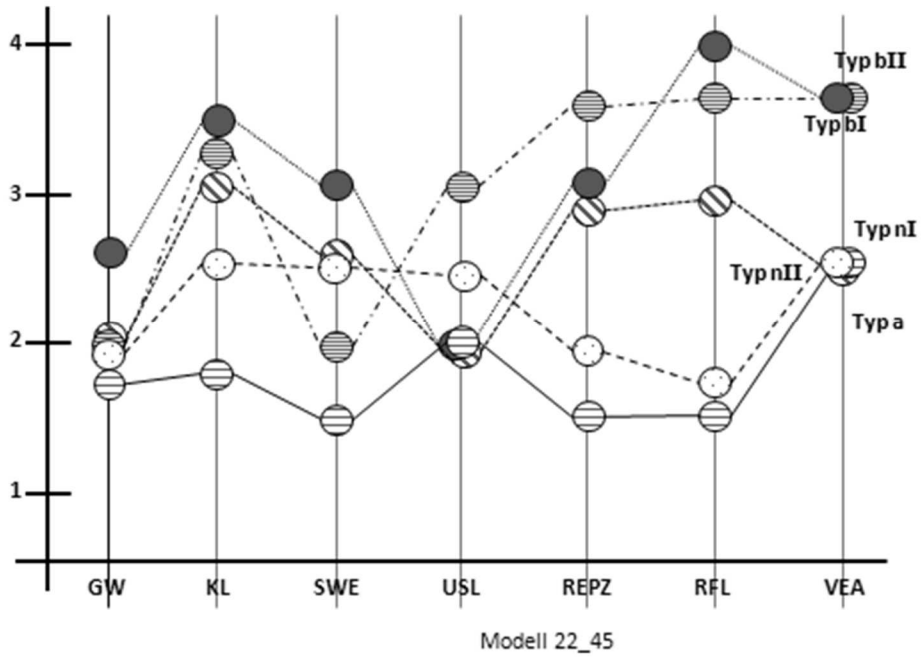


Abbildung 7.19: inhaltliche Darstellung der Klassen zum Modell M22 für fünf Klassen – Klassenmittelwerte abgetragen

Insgesamt erweist damit in der zweiten Kategorie das Modell M22 mit der Vier-Klassenlösung als das beste Modell. Zwar wäre die Fünf-Klassenlösung für das Modell M22 ebenfalls mit den statistischen Kennzahlen vereinbar und die inhaltliche Unterscheidung innerhalb der Klasse „Typ Blockade“ bleibt stimmig, trotzdem ist die weitere Ausdifferenzierung der Klasse bzgl. des Ressourcendefizits im Rahmen des theoretischen Rahmens nicht zwingend notwendig, sodass dem einfacheren Modell Vorrang eingeräumt werden kann. Das Modell M21 hat sich hingegen nicht als sinnvoll realisierbar erwiesen, weil sie weder qualitativ sinnvoll interpretierbare Muster noch geordnete Niveaustufen herausbildeten.

Für die weitergehenden Analysen haben sich damit die Modelle M11a und M22 herauskristallisiert. Auf ihrer Grundlage werden in den folgenden Abschnitt (M11a in 7.3, M22 in 7.4) die deskriptiven Ergebnisse aus (7.1) erneut, aber differenzierter analysiert und dargestellt.

7.3 Eine nach Expertengrad differenzierte Auswertung des Vorbereitungsverhaltens – Expertenklassen für das Unterrichten

Wie in Abschnitt 7.2 dargelegt wurde, können die Klasseneinteilungen des ersten Theoriestrangs am besten für das auf der Datengrundlage der Studie A berechnete Modell M11a_44 [130] interpretiert werden. In diesem Abschnitt wird daher eine Analyse des Vorbereitungsverhaltens der für dieses Modell berechneten vier Klassen (Typ A‘, Typ S‘, Typ B‘ und Typ G‘) vorgenommen. Die Analyse des Vorbereitungsverhaltens orientiert sich dabei an der Darstellung der deskriptiven Ergebnisse aus allen vier Teilstudien. Erneut werden der zeitliche Umfang der Vorbereitung (7.3.1) und die inhaltliche Gestaltung der Vorbereitung (7.3.2) analysiert, anschließend wird ein Blick auf den Gebrauch von Vorbereitungs- und Kompetenzheften geworfen (7.3.3) und das außerschulische Vorbereitungsverhalten wird nach Klassen differenziert dargestellt (7.3.4). In Abschnitt 7.3.5 wird daher nur zwischen „Erfahrung mit VERA8“ und „keine Erfahrung mit VERA8“ unterschieden, um Überschneidungen der Klasseneinteilung des Modells M11a_44 zu untersuchen. In dem Abschnitt wird außerdem wiederum kurz auf die von den befragten Lehrkräften wahrgenommene Bedeutung von VERA8 eingegangen, um das hier beschriebene Bild der Klasseneinteilung zu ergänzen.

7.3.1 Zeitlicher Umfang der Vorbereitung

Für die N=372 klassifizierten Lehrkräfte liegt für n=363 eine Angabe zum für die Vorbereitung auf VERA8 genutzten Stundenvolumen vor. Der Mittelwert über alle klassifizierten Lehrkräfte beträgt M=7.34 Unterrichtsstunden (SE=0.29, SD=5.59, 95% CI=[6.80, 7.95]) und liegt damit umgerechnet bei zwei bis zweieinhalb Schulwochen. Stellt man die durchschnittlich aufgebrachte Unterrichtszeit für die vier Typen des Modells M11a_44 gegenüber, fällt als erstes der hohe Mittelwert der als G‘-Typ‘ klassifizierten Lehrkräfte ins Auge. Er liegt mit M=8.36 Unterrichtsstunden über den Werten für die Lehrkräfte des Typs S‘, M=6.82, t(119.12)=1.728[131], p=.04, r=.17, und denen des Typs B‘, M=6.97, t(134.31)=1.416, p=.08, r=.16. Der Unterschied ist aber nur schwach bzw. nicht signifikant. Lehrkräfte mit hohen eigenen Ressourcen und damit mit besonders guten Voraussetzungen, um guten Unterricht (vgl. Kap. 5) anbieten zu können, bereiteten ihre Klassen durchschnittlich eine Stunde länger auf VERA8 vor als Lehrkräfte, die über weniger der wichtigen Ressourcen verfügen. Dieser Schluss bleibt eingeschränkt auch erhalten, wenn man Lehrkräfte des Typs G‘ und des Typs A‘ bzw. des Typs S‘ und des Typs B‘ jeweils zu einer Klasse zusammenfügt und gegenüberstellt. Auch dann weisen die gemeinsamen Mittelwerte

[130] „_44“ steht dabei für eine Vier-Klassenlösung, die auf Grundlage von vierstufigen Skalen gebildet wurde.
[131] jeweils Welch-Test für unabhängige Stichproben

(M=8.09, SE=.51, SD=6.04, 95% CI=[7.12, 9.14] bzw. M=6.87, SE=.35, SD=5.24, 95% CI=[6.22, 7.57]) einen – wiederum auch nur schwach – signifikanten Unterschied auf, t(267.13)=1.965, p=.03. Das für die Vorbereitung aufgewendete Stundenvolumen war folglich bei den Lehrkräften höher, die allgemein durch ein höheres Arbeitsengagement und Zutrauen in die eigenen Fähigkeiten auffallen.[132] Einschränkend muss allerdings angemerkt werden, dass die Effektstärke dieses Unterschieds mit r=.11 in der gleichen niedrigen Größenordnung anzusiedeln ist wie der Post-it-Effekt (vgl. 7.1.1).

Sichtbar wird dieser wegen der geringen Effektstärke nur als tendenziell zu bezeichnende Unterschied auch noch einmal in der Graphik 7.20. Während in der Klasse TypS‘ & TypB‘ der größte Anteil auf ein bis vier Stunden Vorbereitung entfällt, ist in der Klasse TypA‘ & TypG‘ der Anteil der Lehrkräfte am größten, die fünf bis acht Stunden vorbereiten.

Gleichzeitig ist aber für beide Klassen erkennbar, dass das Mindestniveau des zeitlichen Vorbereitungsumfangs über den akzeptablen ein bis zwei Unterrichtsstunden liegt. Auch in der Klasse TypS‘ & TypB‘ sind es weniger als ein Sechstel, die ihre Klasse nicht länger als zwei Unterrichtsstunden vorbereitet haben und mehr als sechzig Prozent der Lehrkräfte investierten auch hier mehr als eine Woche der Unterrichtszeit.

Tabelle 7.24

Anzahl der für die Vorbereitung aufgewendeten Unterrichtsstunden – differenziert nach Typen M11a_44

Klasse	n	M (SD)	95% CI
A‘	68	7.79 (5.40)	[6.57, 9.00]
S‘	151	6.82 (5.35)	[6.05, 7.74]
B‘	71	6.97 (5.03)	[5.86, 8.20]
G‘	73	8.36 (6.61)	[6.96, 10.00]

[132] Diese pauschale Aussage gilt für das gemessene Arbeitsengagement und für das fachdidaktische Fähigkeitsselbstkonzept. Für die Selbstwirksamkeitserwartung zeigen sich bei differenzierterer Betrachtung nur minimale Unterschiede und für die vierte Klasse wieder ein Anstieg der aufgewendeten Unterrichtszeit.

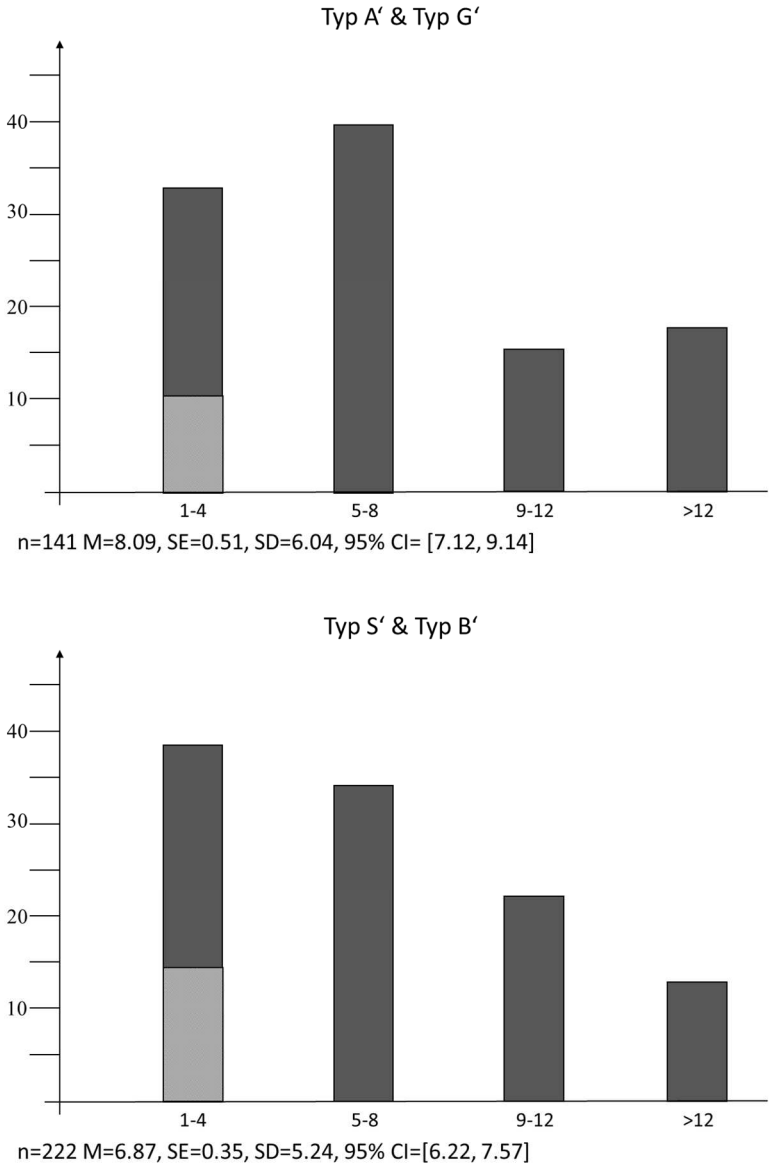


Abbildung 7.20: aufgewendetes Stundenvolumen nach Schulwochen zusammengefasst – Gegenüberstellung der Typen M11a_44 S & B’ und A’& G’ (Verteilung der Lehrkräfte in Prozent – heller Anteil: Lehrkräfte mit max. zweistündiger Vorbereitung)

7.3.2 Gestaltung der Vorbereitungszeit: Familiarity Approach, Content Approach und Test-Wiseness-Strategien

Wiederum als zweiter Indikator für die Intensität der Vorbereitung soll die Anzahl der durchgeführten Maßnahmen dienen. Dazu sollen wieder zuerst die acht Maßnahmen betrachtet werden, mit denen Familiarity Approach (FA) durchgeführt werden kann, sowie der spezielle Umgang mit der Aufgabenstellung, um die vier Klassen zu vergleichen. Hier zeigt sich allein über die Kennzahlen der Maßnahmen kein bemerkbarer Unterschied einer Klasse zu den drei anderen. In drei der vier Klassen ist der häufigste Wert die Durchführung von zwei Maßnahmen im Sinne eines FA (Typen A', S', & G'), dreimal haben über die Hälfte der Lehrkräfte maximal zwei Maßnahmen durchgeführt (Typen A', S' & B'). Obwohl die Klasse Typ S' mit sechs Maßnahmen das größte Maximum besitzt und die Klasse Typ B' als einzige den Modalwert drei aufweist, sind in den beiden anderen Klassen jeweils mehr Lehrkräfte vorhanden, die vier oder mehr Maßnahmen durchgeführt haben.

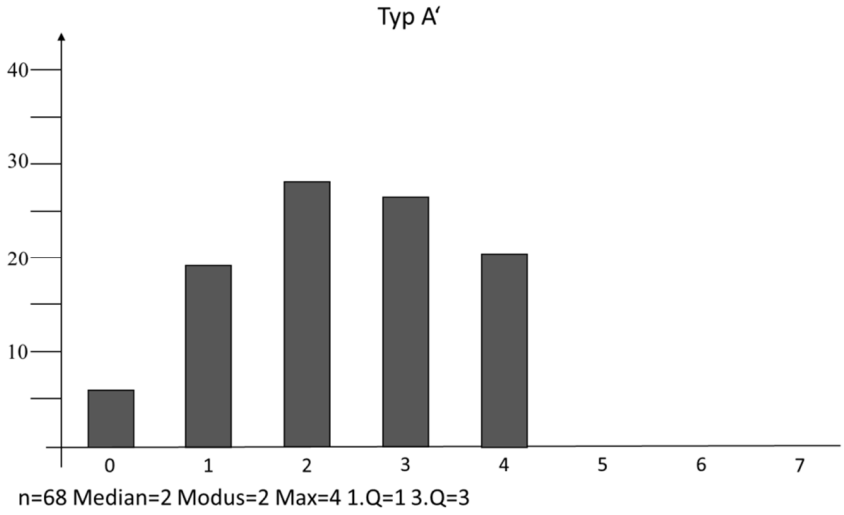


Abbildung 7.21: Darstellung der Nutzungsvariabilität von FA-Maßnahmen – Typ A' M11a_44 (Verteilung der Lehrkräfte in Prozent)

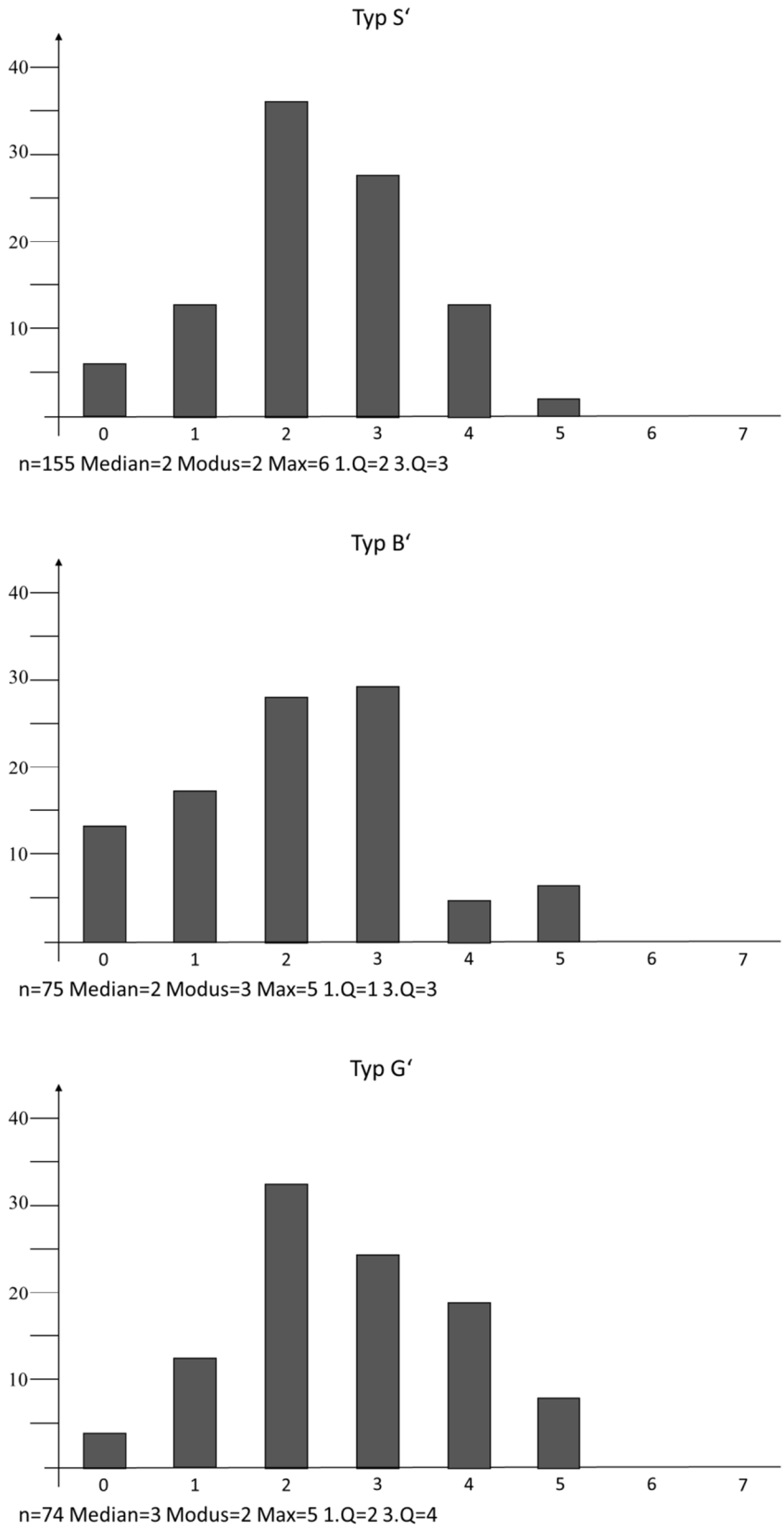
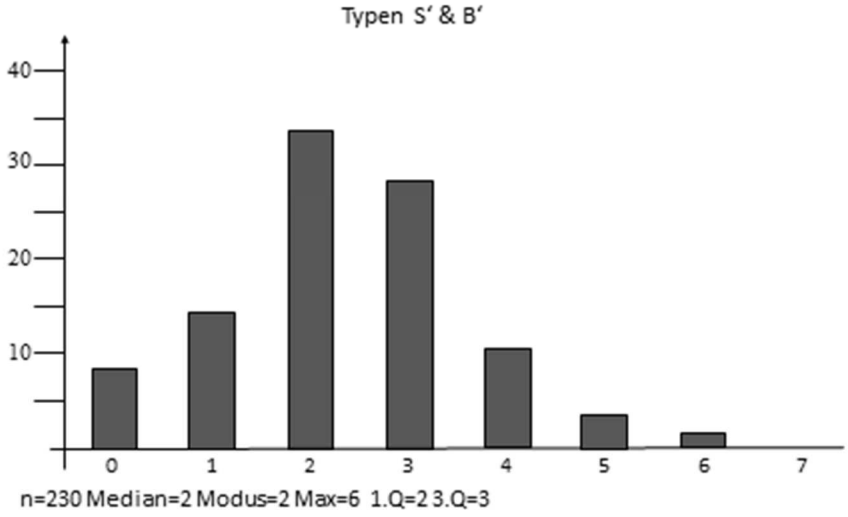


Abbildung 7.22: Darstellung der Nutzungsvariabilität von FA-Maßnahmen – Typen S‘, B‘ & G‘ M11a_44 (Verteilung der Lehrkräfte in Prozent)

Dies bestätigt erneut der Blick auf die zusammengefassten Klassen Typ S‘ & Typ B‘ bzw. Typ A‘ & Typ G‘. Wie man in den Graphiken 7.21 bis 7.24 erkennen kann, unterscheidet sich die Verteilung der Maßnahmenhäufigkeit für A’ & G’ für Säulen zu zwei, drei und vier Maßnahmen weniger deutlich als für die beiden anderen Muster. Die Vorbereitung durch Lehrkräfte der Typen A’ und G’ war folglich also leicht variantenreicher als die Vorbereitung durch Lehrkräfte, die den Typen S’ und B’ zugeordnet wurden.



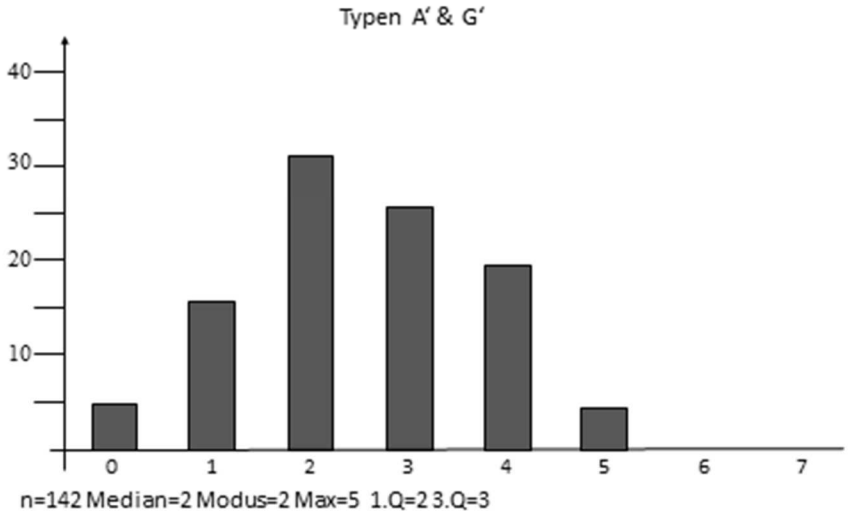


Abbildung 7.25: Darstellung der Nutzungsvariabilität von FA-Maßnahmen unterschieden nach Lehrkräften der Typen M11a_44 A’ & G’ und S’ & B’ (Verteilung der Lehrkräfte in Prozent)

Obwohl dieser Unterschied gleichermaßen auf den Lehrkräften des Typs A' wie des Typs G' beruht, zeigen sich vor allem für den Typ G' interessante Unterschiede zu den anderen beiden Klassen, wenn die Nutzung der einzelnen Maßnahmen differenziert betrachtet wird. Hierbei hilft nachfolgende Tabelle 7.25. Entsprechend der globalen Betrachtung in (7.1) wurde als häufigste Maßnahme (a) das Übenlassen mit alten LSE-Aufgaben von allen vier Typen genannt. Lehrkräfte des Typs G' nannten die Durchführung dieser Maßnahme allerdings (schwach) signifikant ($\chi^2[1]=4.872$, p=.03, w=.11) häufiger (91.9%) als die Lehrkräfte der drei anderen Typen (jeweils um achtzig Prozent). Auch die Maßnahmen (b), (c) und (d) wurden von Lehrkräften des Typs G' häufiger durchgeführt. Der größte Unterschied liegt dabei aber direkt zwischen den Lehrkräften des Typs G' und den Lehrkräften des Typs B'. Soweit sich dies in dem eingesetzten Fragebogen abbilden lässt, scheint es einen Unterschied in der Nutzung der Aufgaben gegeben zu haben. Lehrkräfte des Typs B' nutzten zwar ebenfalls – wenn auch weniger häufig – Aufgaben aus zentralen Lernstandserhebungen im Unterricht, der Aufgabentypus fand aber seltener Niederschlag in den anderen Teilen der Leistungsmessung. Überraschend sind die hohen Werte der Lehrkräfte des Typs S'. Betrachtet man (c) und (d) gemeinsam als Indikator, ob durch VERA8 ein anderer Aufgabentyp implementiert wird, steigt der Wert für Lehrkräften dieses Typs noch einmal auf 36.1% (Typ G': 40.5%, Typ A': 32.3%, Typ B': unverändert). Ebenfalls auffällig ist der Unterschied bei der Maßnahme (g), Schüler auf die offizielle Internetseite hinzuweisen, zwischen Lehrkräften des Typs G' (64.9%) und des Typs B' (48.0%) ($\chi^2[1]=4.308$, p=.04, w=.17).

Tabelle 7.25

Welche der folgenden Optionen haben Sie in dieser Klasse als unmittelbare Vorbereitung auf die LSE genutzt? – Familiarity Approach-Maßnahmen differenziert nach Typen M11a_44

Maßnahme	Typ A' (n=68) absolut (%)	Typ S' (n=155) absolut (%)	Typ B' (n=75) Absolut (%)	Typ G' (n=74) Absolut (%)
(a) mit Testaufgaben früherer LSE üben lassen	56 (82.4%)	128 (81.9%)	59 (78.7%)	68 (91.9%)
(b) alte Testaufgaben zur Verfügung gestellt	44 (64.7%)	102 (65.8%)	43 (57.3%)	51 (68.9%)
(c) zu LSE ähnliche Aufgaben in vorherige Klassenarbeiten eingebaut	21 (30.9%)	49 (31.6%)	18 (24.0%)	29 (39.2%)

Maßnahme	Typ A' (n=68) absolut (%)	Typ S' (n=155) absolut (%)	Typ B' (n=75) Absolut (%)	Typ G' (n=74) Absolut (%)
(d) alte LSE-Aufgaben in Klassenarbeiten eingebaut	4 (5.9%)	17 (11.0%)	3 (4.0%)	14 (18.9%)
(e) Beispielaufgaben von Homepage lösen lassen	17 (25.0%)	38 (24.5%)	21 (28.0%)	18 (24.6%)
Familiarity Approach mit Aufgaben	64 (94.1%)	144 (92.9%)	65 (88.0%)	70 (94.6%)
(f) in der Klasse gemeinsam die offizielle Internetseite besucht	4 (5.9%)	14 (9.0%)	2 (2.7%)	4 (5.4%)
(g) Schüler auf offizielle Internetseite hingewiesen	41 (60.3%)	84 (54.2%)	36 (48.0%)	48 (64.9%)
(h) Testsituation simuliert	15 (22.1%)	33 (21.3%)	16 (21.3%)	13 (17.6%)

Auch für die Klasseneinteilung nach M11a_44 kann betrachtet werden, inwieweit im Sinne eines Familiarity Approach bestimmte Bestandteile der Aufgaben Teil der Vorbereitung waren. Tabelle 7.26 stellt dies für die vier Typen jeweils dar. Auffällig ist ein grundsätzlich größeres Zurückhalten von Lehrkräften des Typs B', die Aufgabenstellungen überhaupt zu thematisieren. Ungefähr jeweils dreißig Prozent der Lehrkräfte dieses Typs haben nicht angesprochen, wie man die Aufgabenstellung richtig versteht, wie man ihr die richtigen Informationen entnimmt oder wie man bei den speziellen Antwortformaten richtig antwortet. 18.3% jener Lehrkräfte haben sogar weder (p) noch (q) noch (r) thematisiert. Im Gegensatz dazu haben Lehrkräfte der drei anderen Typen die Themen (p) und (r) häufiger thematisiert und in mehr als der Hälfte der Fälle auch mehrfach angesprochen. Besonders deutlich ist der Unterschied zwischen den Lehrkräften des Typs B' und jenen des Typs A' ($\chi^2_p[2]=4.699$, p=.10, w=.19; $\chi^2_r[2]=8.960$, p=.01 w=.25). Lehrkräfte des Typs A' haben demgegenüber auch nur zu 3.2% gar keines der drei Themen angesprochen. Sie investierten dabei besonders deutlich Unterrichtszeit in die Vermittlung, wie man die Aufgabenstellung richtig versteht, während der (noch deutlichere) Unterschied bei (r) vorwiegend darauf zurückzuführen ist, dass mehr Lehrkräfte aus dieser Klasse es überhaupt sinnvoll fanden, die

Antwortformate anzusprechen. Es zeigt sich aber auch, dass selbst für 30% der Lehrkräfte des Typs G' die Entnahme der relevanten Informationen kein ausreichend wichtiges Thema war.

Nach den FA-Maßnahmen werden als nächstes Unterschiede in der Umsetzung der vorgegebenen Maßnahmen betrachtet, die unter Content Approach (CA) fallen. Wie bereits in (7.1.3) beschrieben, wurde im Fragebogen zwischen Inhalts- und Prozesskompetenzen unterschieden und es wurde abgefragt, ob diese jeweils im Unterricht thematisiert wurden und bzw. oder den Schülern als Übungsbereiche empfohlen wurden. In Tab. 7.27 sind anders als in Tab. 7.4 unter (j*) und (m*) nur diejenigen Lehrkräfte angegeben, die *ausschließlich* ihren Schülern empfohlen haben, die Inhalts- und Prozesskompetenzen zu wiederholen, diese aber nicht selbst im Unterricht behandelt haben.

Tabelle 7.26

Welche der folgenden Themen haben Sie im Unterricht angesprochen? – differenziert nach Typen M11a_44

	Typ A‘ (n=68)			Typ S‘ (n=155)			Typ B‘ (n=75)			Typ G‘ (n=74)		
Maßnahme	mehrfach (%)	Einmal (%)	gar nicht (%)	mehrfach (%)	einmal (%)	gar nicht (%)	mehrfach (%)	einmal (%)	gar nicht (%)	mehrfach (%)	einmal (%)	gar nicht (%)
(p) wie man die Aufgabenstellung der LSE-Aufgaben richtig versteht	38 (59.4%)	17 (25.9%)	9 (13.2%)	80 (53.3%)	39 (26.0%)	31 (20.7%)	35 (49.3%)	15 (21.1%)	21 (29.6%)	41 (57.7%)	15 (19.7%)	16 (22.5%)
(q) wie man den Aufgabenstellungen der LSE-Aufgaben die wichtigen Informationen entnimmt	34 (55.7%)	12 (19.7%)	15 (24.6%)	79 (53.0%)	31 (20.8%)	39 (26.2%)	35 (48.6%)	16 (22.2%)	21 (29.2%)	36 (51.4%)	13 (18.6%)	21 (30.0%)
(r) wie man in Hinblick auf die speziellen Antwortformate richtig antwortet	33 (50.8%)	24 (36.9%)	8 (12.3%)	83 (56.5%)	39 (26.5%)	25 (17.0%)	35 (47.9%)	15 (20.5%)	23 (31.5%)	37 (51.4%)	21 (29.2%)	14 (19.4%)

Anmerkung: Die oben zu den Typen gegebene Anzahl bezieht sich immer auf die Anzahl der Lehrkräfte, zu denen mindestens zu einer der Optionen eine Angabe vorlag. Es können sich daher für die einzelnen Optionen Abweichungen in der Summe ergeben.

Der Vergleich der vier Klassen zeigt wieder, dass sich die Typen A' & G' und die Typen S' & B' unterscheiden. Lehrkräfte der Typen A' oder G' haben häufiger ihren Unterricht genutzt, um vor VERA8 noch einmal alle Inhaltsbereiche und alle oder zumindest eine Prozesskompetenz zu wiederholen. Lehrkräfte der Typen A' und G' haben (schwach signifikant) häufiger alle Inhaltsbereiche noch einmal im Unterricht wiederholt als Lehrkräfte der beiden anderen Typen ($\chi^2_i[1]=4.179$, p=.041 w=.11). Umgekehrt haben Lehrkräfte der Typen S' und B' minimal häufiger diese Wiederholung ihren Schülerinnen und Schülern empfohlen, aber nicht selbst Unterrichtszeit dafür verwendet. Lehrkräfte der Typen S' und B' haben aber zumindest in kleiner Zahl *einzelne* Inhaltsbereiche wiederholt. In der Klasse Typ B' gilt letzteres sogar für so viele Lehrkräfte, dass sie insgesamt gesehen häufiger eine inhaltliche Wiederholung durchführten als die Lehrkräfte der Typen A' & G' (vgl. Tab. 7.27, letzte Zeile).

Ähnlich verhält es sich im Umgang mit den Prozesskompetenzen. Die allgemeinen deskriptiven Befunde aus allen vier Teilstudien haben schon gezeigt, dass die Prozesskompetenzen eine geringere Bedeutung im Vergleich zu den Inhaltskompetenzen zu besitzen schienen. Hierbei sind die Unterschiede zwischen den Lehrkräften mit hohem Arbeitsengagement und positiven Kompetenz- und Kontrollüberzeugungen und Lehrkräften, die hierin Defizite aufweisen, für unterrichtliche Wiederholungsphasen noch deutlicher und hoch signifikant ($\chi^2_k[1]=10.380$, p=.001 w=.17). Besonders Lehrkräfte des Typs S' zeichneten sich dadurch aus, dass sie überdurchschnittlich häufig eine Wiederholung der Inhalts- und Prozesskompetenzen ihren Schülern und Schülerinnen nur empfahlen statt selbst durchzuführen. Auch dies könnte als Indikator für ein Zutrauen in die selbstregulativen Fähigkeiten der Schülerinnen und Schüler gewertet werden. Im Unterschied zum Hinweis auf die offizielle Internetseite (vgl. 7.3.2) sind hier in der Tat nur Lehrkräfte ausgewiesen, die die Themen nicht flankierend im Unterricht wiederholten. Der geringere Stellenwert der Prozesskompetenzen bei Lehrkräften der Typen S' und B' zeigt sich aber auch, wenn man diejenigen Lehrkräfte berücksichtigt, die das Wiederholen jener Kompetenzen nur ihren Schülern und Schülerinnen aufgetragen haben und evtl. nur eine Prozesskompetenz im Unterricht wiederholt haben (χ^2_l [1=10.380, p=.004 w=.15).

Tabelle 7.27

Welche der folgenden Optionen haben Sie in dieser Klasse als unmittelbare Vorbereitung auf die LSE genutzt? – CA-Maßnahmen differenziert nach Typen M11a_44

Maßnahme	Typ A' (n=68) absolut (%)	Typ S' (n=155) absolut (%)	Typ B' (n=75) Absolut (%)	Typ G' (n=74) Absolut (%)
(i) *alle* Inhaltsbereiche wiederholt	34 (50.0%)	59 (38.1%)	28 (37.3%)	35 (47.3%)
(j)* *nur* Empfohlen, alle Inhaltsbereiche zu wiederholen	15 (22.1%)	40 (25.8%)	22 (29.3%)	18 (24.3%)
alle Inhaltsbereiche wiederholt und zus. Empfohlen, alle zu wiederholen	18 (26.4%)	34 (21.9%)	14 (18.7%)	15 (20.3%)
(i) und/oder (j)	49 (72.1%)	99 (63.9%)	50 (66.7%)	53 (71.6%)
(k) *alle* Prozessbereiche wiederholt	22 (32.4%)	27 (17.4%)	18 (24.0%)	27 (36.5%)
(l)* *nur* empfohlen, alle Prozesskompetenzen zu wiederholen	6 (8.8%)	27 (17.4%)	8 (10.7%)	8 (10.8%)
alle Prozesskompetenzen wiederholt und zus. Empfohlen, alle zu wiederholen	10 (14.7%)	9 (5.8%)	9 (12.0%)	11 (14.9%)
(m) *eine* Prozesskompetenz besonders üben lassen	11 (16.2%)	13 (8.4%)	11 (14.7%)	11 (14.9%)
(k)+(l)+(m)	37 (54.4%)	61 (39.4%)	30 (40.0%)	41 (55.4%)
Lehrkräfte, die mindestens eine der vorformulierten CA-Maßnahmen umsetzten	55 (80.9%)	105 (67.7%)	55 (73.3%)	58 (78.4%)
Lehrkräfte, die mindestens eine der vorformulierten oder der freiformulierten CA-Maßnahmen umsetzten	56 (82.4%)	112 (72.3%)	61 (81.3%)	60 (81.1%)

Als letztes Element zur Gestaltung der Vorbereitung wird auch für die vier Klassen des Modells M11a_44 differenziert, welche Tipps & Tricks im Vorfeld zu VERA8 von den Lehrkräften angesprochen wurden. Die Ergebnisse dazu sind in Tab. 7.28 dargestellt. Es sind für einzelne Test-Wiseness-Strategien (TWS) leichte Unterschiede zwischen den Klassen zu erkennen, es gibt aber keine grundsätzliche Tendenz. Insbesondere gibt es daher auch keinen Unterschied bei der Vermittlung der Strategien, die darauf abzielen, die Chancen zu erhöhen, tatsächlich zu der richtigen Lösung im Sinne der Aufgabenintention zu gelangen (TWS2, TWS3 und TWS4). Nur für die TWS5 existiert ein hochsignifikanter Unterschied

zwischen A' & G' und S' & B' mit 36.4% zu 20.5%, $\chi^2(1)=10.663$, p=.001, w=.17. Etwas höher ist unter den Lehrkräften der Typen A' und G' auch die Anzahl, die WST8 vermittelt haben, wobei der Unterschied nicht signifikant und die Effektstärken minimal sind (31.5% zu 23.2%, $\chi^2[1]=2.793$, p=.10, w=.09).

Tabelle 7.28

Welche der folgenden Strategien haben Sie im Unterricht angesprochen? – differenziert nach Typen M11a_44

	Typ A'		Typ S'		Typ B'		Typ G'	
	Ja	Nein	Ja	Nein	Ja	Nein	Ja	Nein
Test-Wiseness-Strategie	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
(TWS1)…sich nicht zu lang an einer Frage aufzuhalten	44 (71.0%)	18 (29.0%)	115 (75.7%)	37 (24.3%)	55 (74.3%)	19 (25.7%)	52 (70.3%)	18 (25.7%)
(TWS2)…sich mit den Antworten vertraut zu machen.	38 (61.3)	24 (38.7%)	102 (68.0%)	48 (32.0%)	48 (64.9)	26 (35.1%)	50 (70.4%)	21 (29.6%)
(TWS3)…alle Antworten in Betracht zu ziehen, bevor man sich entscheidet.	44 (71.0%)	18 (29.0%)	95 (62.1%)	58 (37.9%)	41 (55.4%)	33 (44.6%)	47 (67.1%)	23 (32.9%)
(TWS4)…die Instruktionen und Fragen genau zu lesen.	64 (95.5%)	3 (4.5%)	140 (90.9%)	14 (9.1)	70 (93.3%)	5 (6.7%)	67 (91.8%)	6 (8.2%)
(TWS5)…bei Multiple-Choice-Fragen zu raten, wenn man die Antwort nicht weiß.	18 (30.0%)	42 (70.0%)	31 (20.5%)	120 (79.5%)	15 (20.5%)	58 (79.5%)	29 (42.0%)	40 (58.0%)
(TWS6)…zuerst die Fragen zu beantworten, bei denen man sich sicher fühlt.	46 (71.9%)	18 (28.1%)	113 (74.3%)	39 (25.7%)	51 (68.9%)	23 (31.1%)	50 (70.4%)	21 (29.6%)
(TWS7)…sich spontane Einfälle zu notieren.	13 (22.4%)	45 (77.6%)	18 (12.5%)	126 (87.5%)	10 (13.7%)	63 (86.3%)	9 (13.2%)	59 (86.8%)
(TWS8)…auf grammatikalische Einschränkungen der möglichen Antworten zu achten.	22 (37.3%)	37 (62.7%)	33 (22.0%)	117 (78.0%)	19 (26.0%)	54 (74.0%)	18 (26.5%)	50 (73.5%)
(TWS9)…im Zweifel die erste Idee zu wählen, weil dies meist das Beste ist.	2 (3.4%)	57 (96.6%)	6 (4.1%)	141 (95.9%)	0	73 (100.0 %)	5 (7.5%)	62 (92.5%)

7.3.3 Nutzung von Vorbereitungs- und Kompetenzheften

Als nächstes sollen mögliche Unterschiede zwischen den vier Klassen des Modells M11a_44 bezüglich der Nutzung von Vorbereitungs- und Kompetenzheften analysiert werden. Zur Erinnerung sei noch einmal darauf hingewiesen, dass beide Heftarten als verschiedene Indikatoren interpretiert werden können: Vorbereitungshefte als Ausdruck einer input-orientierten Steuerung und Kompetenzhefte als Möglichkeit einer output-orientierten Steuerung. Die grundsätzliche Tendenz zeigte eine häufigere Nutzung von Vorbereitungsheften (vgl. auch Kap. 8).

Der Vergleich der vier Klassen offenbart dieses Ergebnis auf den ersten Blick wiederum für alle vier Klassen fast gleichermaßen (vgl. Tab. 7.29). In allen vier Klassen haben deutlich mehr Lehrkräfte Vorbereitungshefte eingesetzt als Kompetenzhefte genutzt. Nur eine minimale Tendenz lässt sich darin ausmachen, wie sich das Verhältnis von eingesetzten Kompetenzheften zum generellen Einsatz von Heften darstellt. Bei diesem Verhältnis wurden Kompetenzhefte von 32.4% der Lehrkräfte des Typs A' eingesetzt, unter Lehrkräften des Typs S' waren es 32.3%, bei Typ B' 37.2% und 42.9% der entsprechenden Lehrkräfte des Typs G' nutzten diese Hefte. Lehrkräfte des Typs G' scheinen also eher als andere Lehrkräfte zu einer Nutzung zu tendieren. Der Unterschied war allerdings nicht signifikant ($\chi^2[1]=1.288$, $p=.256$). Außerdem sollte man an dieser Stelle nicht übersehen, dass insgesamt nur etwas mehr als die Hälfte der Lehrkräfte überhaupt eine Art von Hefte in der Vorbereitung einsetzte.

Aufschlussreicher ist eine genauere Analyse der Nutzung von Vorbereitungsheften. Während Lehrkräfte der Typen A', B' und G' zu jeweils mehr als einem Drittel die Hefte nutzten, um alle Inhaltsbereiche zu wiederholen, und zumindest ein Viertel von ihnen dies auch für die Wiederholung der Prozesskompetenzen angab, ist der Anteil an Lehrkräften in der Klasse des Typs S' beide Male geringer und der Anteil derjenigen Lehrkräfte, die die Prozesskompetenzen gar nicht beabsichtigten zu wiederholen, besonders hoch.

Trotz der insgesamt geringen Fallzahlen der hier gegenüber gestellten Gruppen ist der Unterschied sowohl für die Inhalts- als auch für die Prozesskompetenzen signifikant mit $\chi^2(2)=12.661$, $p=.002$, $w=.26$ bzw. $\chi^2(2)=8.999$, $p=.011$, $w=.22$. Lehrkräfte des Typs S' nutzten Vorbereitungshefte folglich weniger zu einer umfassenden Wiederholung, sondern vielmehr zu einer selektiv[133]. Gleichzeitig zeigt eine genauere Analyse derjenigen Lehrkräfte, die keine Vorbereitungshefte nutzten, in Bezug auf die Wiederholung der Prozesskompetenzen: Nur 10.3% der Lehrkräfte wiederholten zumindest einige Prozesskompetenzen, ohne dazu

[133] „Selektiv" ist hier nicht mit „zielgerichtet" gleichzusetzen. Die Gründe, warum von Lehrkräften des Typs S nur manche Inhalts- und Prozesskompetenzen wiederholt wurden, wurden in dieser Studie nicht erfasst. Es ist nicht unwahrscheinlich, dass zu wenig Zeit zur Verfügung stand. Möglicherweise lag es sogar in der Absicht dieser Lehrkräfte, alle Kompetenzen zu wiederholen.

Vorbereitungshefte zu nutzen. Bei den anderen drei Typen waren es 40.0% (Typ A'), 28.6% (Typ B') und 34.3% (Typ G').

Tabelle 7.29

Einsatz von Vorbereitungs- und Kompetenzheften im Unterricht - differenziert nach Typen M11a_44

Maßnahme	Typ A' (n=68) Absolut (%)	Typ S' (n=155) absolut (%)	Typ B' (n=75) absolut (%)	Typ G' (n=74) Absolut (%)
Vorbereitungshefte eingesetzt	33 (48.5%)	87 (56.1%)	40 (53.3%)	39 (52.7%)
Kompetenzhefte eingesetzt	12 (17.6%)	30 (19.4%)	16 (21.3%)	18 (24.3%)
Beide Arten eingesetzt	8 (11.8%)	24 (15.5%)	13 (17.3%)	14 (18.9%)
Lehrkräfte, die mindestens eine Heftart eingesetzt haben	37 (54.4%)	93 (60.0%)	43 (57.3%)	42 (56.8%)

Überhaupt unterscheiden sich Lehrkräfte, die Vorbereitungshefte nutzten, von denjenigen, die diese nicht nutzen, in den einzelnen Klassen des Modells M11a_44. Für Lehrkräfte der Typen A', S' und G' zeigte sich bei Nutzung von Vorbereitungsheften im Unterricht eine Zunahme der Vorbereitungsvariabilität von FA-Maßnahmen. In allen drei Klassen ändert sich der Median von zwei (Vorbereitungshefte nicht eingesetzt) auf drei (Vorbereitungshefte eingesetzt). Für die Klasse des Typs A' gilt sogar, dass Lehrkräfte hieraus ohne Vorbereitungshefte am häufigsten nur eine der sieben Maßnahmen umsetzten, während es bei Lehrkräften mit Vorbereitungsheften drei Maßnahmen waren.

Tabelle 7.30

Wiederholung von Inhalts- und Prozessbereichen mit Vorbereitungsheften – differenziert nach Typen M11a_44

	Typ A'	Typ S'	Typ B'	Typ G'
	(n=33)	(n=87)	(n=40)	(n=39)
	absolut	absolut	Absolut	Absolut
Maßnahme	(%)	(%)	(%)	(%)
alle Inhaltsbereiche wiederholt	13 (39.4%)	14 (16.3%)	13 (34.2%)	14 (38.8%)
einzelne Inhaltsbereiche wiederholt	20 (60.6%)	71 (82.6%)	23 (60.5%)	20 (55.6%)
	(n=33)	(n=83)	(n=40)	(n=39)
alle Prozessbereiche wiederholt	9 (28.1%)	9 (10.3%)	10 (25.6%)	10 (27.0%)
einzelne Prozessbereiche wiederholt	21 (65.6%)	64 (77.1%)	27 (69.2%)	25 (67.6%)

Anmerkung: Die Anteile ergänzen sich zu 100% mit denjenigen, die „weder noch" angaben.

Tabelle 7.31

Vorbereitungsvariabilität und Einsatz von Vorbereitungsheften - differenziert nach Typen M11a_44

	Typ A' (n=68)				Typ S' (n=155)				Typ B' (n=75)				Typ G' (n=74)			
Heftnutzung	n	Mdn	X_D	∑	n	Mdn	X_D	∑	n	Mdn	X_D	∑	n	Mdn	X_D	∑
Hefte eingesetzt	33	3	3	87	87	3	2	22	40	2	2	83	39	3	2	112
Hefte nicht eingesetzt	35	2	1	74	68	2	2	152	35	2	3	79	35	2	2	83

Auch ein analoger Vergleich der aufgewendeten Vorbereitungszeit offenbart sogar für alle vier Modellklassen eine Zunahme des Stundenvolumens. Die Ausgangsbasis stellen ca. sechs Schulstunden Vorbereitungszeit dar, die Lehrkräfte durchschnittlich aufwenden, um ihre Schülerinnen und Schüler ohne Zuhilfenahme von Vorbereitungsheften speziell auf VERA8 vorzubereiten. Hier verhält sich der Typen-Vergleich aber teilweise konträr zum Befund bezüglich der FA-Maßnahmen. Während die Lehrkräfte aus den Teilklassen mit Nutzung von Vorbereitungsheften in den anderen drei Modellklassen drei bis vier Stunden mehr vorbereiten und sich die Teilklassen dort signifikant unterscheiden (Typ A‘: t[65.14]=2.385, p=0.01 r.39; Typ B‘: t[58.71]=2.496, p=0.01, r.38 und Typ G‘: t[62.62]=2.934, p<0.01, r.41), zeigt sich für die Modellklasse Typ S’ nur ein nicht signifikanter Unterschied von ca. einer Stunde.

Lehrkräfte des Typs S’ steigerten somit also die Qualität ihrer Vorbereitung im Sinne einer variantenreicheren Vorbereitung, sie erhöhten aber nicht gleichzeitig das Stundenvolumen der Vorbereitung. In diesem Zusammenhang ist sicherlich auch die Beobachtung zu sehen, dass Lehrkräfte dieses Typs mit Vorbereitungsheften häufiger als andere Lehrkräfte nur einzelne Inhalts- und Prozesskompetenzen wiederholt und geübt haben. Umgekehrt steigerten Lehrkräfte des Typs B’ zwar das Stundenvolumen, welches sie für die Vorbereitung auf VERA8 aufwendeten, wenn sie Vorbereitungshefte einsetzten, dabei nutzten sie aber nicht mehr FA-Mahnahmen und bereiteten vorwiegend auf die Inhaltskompetenzen vor, weniger auf die Prozesskompetenzen.

Tabelle 7.32

Vorbereitungsumfang und Einsatz von Vorbereitungsheften - differenziert nach Typen M11a_44

	Typ A' (n=68)			Typ S' (n=155)			Typ B' (n=75)			Typ G' (n=74)		
Heftnutzung	n	M (SD)	95% CI	n	M (SD)	95% CI	n	M (SD)	95% CI	N	M (SD)	95% CI
Hefte eingesetzt	33	9.39 (5.36)	[7.55, 11.72]	87	7.44 (4.76)	[6.47, 8.43]	40	8.32 (5.88)	[6.27, 10.27]	39	10.40 (7.50)	[8.26, 12.95]
Hefte nicht eingesetzt	35	6.29 (5.07)	[4.80, 7.91]	68	6.03 (5.98)	[4.80, 7.64]	35	5.50 (3.42)	[4.32, 6.65]	35	6.43 (4.67)	[4.74, 7.80]

Tabelle 7.33

Vorbereitungszeit nach Erfahrung und eingeschätzter Veränderung der Intensität - differenziert nach Typen M11a_44

	Typ A'			Typ S'			Typ B'			Typ G'		
Erfahrung	n	M (SD)	95% CI	n	M (SD)	95% CI	n	M (SD)	95% CI	n	M (SD)	95% CI
Bisher keine Erfahrung mit VERA8	20	8.90 (6.17)	[6.55, 11.5]	64	5.92 (3.86)	[5.06, 6.81]	26	5.77 (4.40)	[4.15, 7.54]	17	7.47 (6.05)	[4.70, 10.41]
Intensität gleichbleibend	26	7.58 (5.12)	[5.77, 9.73]	53	8.19 (7.25)	[6.40, 10.30]	30	8.67 (5.85)	[6.70, 10.83]	33	9.30 (8.07)	[6.91, 12.25]

7.3.4 Außerunterrichtliche Vorbereitung

Die bisherigen Analysen haben bezüglich der Vorbereitung auf VERA8 durch die Lehrkräfte innerhalb der Unterrichtszeit wenig statistisch bedeutsame Unterschiede zwischen den vier Klassen des Modells M11a_44 offenbart, lassen aber grundsätzlich Tendenzen erkennen. Dazu gehört, dass Lehrkräfte der Typen S' und B' weniger Stunden für die Vorbereitung nutzten als Lehrkräfte der Typen A' und G'. Genauso investierten sie weniger in ein Vertrautwerden mit VERA8 (Typ B') und in die Prozesskompetenzen (Typ S'). Inhaltlich und bezogen auf den zeitlichen Umfang erlebten deren Schüler folglich ein Defizit. Nicht auszuschließen ist, dass ihre Schülerinnen und Schüler insgesamt auch weniger Zutrauen in die Fähigkeiten jener Lehrerinnen und Lehrer hatten, weil sich ihre Defizite innerhalb der personenbezogenen Ressourcen auf ihre Lehrtätigkeit ausgewirkt haben sollten. Gleichzeitig haben leicht mehr Lehrkräfte der Typen S' und B' Vorbereitungshefte im Unterricht genutzt. Dadurch sollte es zu Unterschieden im außerschulischen Vorbereitungsverhalten der Schüler gekommen sein.

Tabelle 7.34

außerunterrichtliche Übungsphasen in der Wahrnehmung der Lehrkräfte – differenziert nach Typen M11a_44

	Typ A'		Typ S'		Typ B'		Typ G'	
	Ja	Nein	Ja	Nein	Ja	Nein	Ja	Nein
	absolut	absolut	absolut	absolut	absolut	absolut	absolut	Absolut
Maßnahme	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Haben Sie der Klasse empfohlen, sich speziell auf VERA8 vorzubereiten?	44 (66.7%)	22 (33.3%)	100 (64.9%)	54 (35.1%)	47 (62.7%)	28 (37.3%)	45 (62.5%)	27 (37.5)
Haben die Schüler für VERA8 außerhalb des Unterrichts besonders geübt?	31 (48.4%)	33 (51.6%)	94 (65.3%)	54 (34.7%)	37 (55.2%)	30 (44.8%)	39 (54.9%)	32 (45.1%)

In der Tat zeigt sich für Lehrkräfte des Typs S' ein leichter Unterschied bezüglich der außerunterrichtlichen Vorbereitung der Schüler auf VERA8. Nach Wahrnehmung der

Lehrkräfte haben sich in über 60% der Fälle die Schüler und Schülerinnen außerhalb des Unterrichts auf VERA8 zusätzlich vorbereitet. Bei den übrigen klassifizierten Lehrkräften sind es zwar auch immerhin knapp 53.0%, der Unterschied ist aber zumindest schwach signifikant ($\chi^2[1]=3.884$, $p<.05$, $w=.11$). Hingegen zeigen sich keine Unterschiede bei der Frage, ob diese Lehrkräfte die außerschulische Vorbereitung auch selbst empfohlen haben. Im Vergleich der Antworten zu diesen Fragen wird deutlich, dass sich nur bei Lehrkräfte des Typs S' die Verteilung der Antworten gleicht, während unter den Lehrkräften der anderen drei Typen mehr Lehrkräfte eine außerschulische Vorbereitung empfohlen haben, aber nicht annahmen, ihre Schüler seien diesem Rat gefolgt. Für Lehrkräfte des Typs B' zeigt sich nicht derart deutlich, dass Schülerinnen und Schüler selbstständig zusätzlich für VERA8 geübt haben. Trotzdem nahmen aber auch unter diesen Lehrkräften mehr als die Hälfte dieses von ihren Schülerinnen und Schülern an.

7.3.5 Veränderung der Vorbereitungsintensität und Bewertung von VERA8

In Abschnitt 7.1.5 wurde für alle vier Teilstudien zwischen Lehrkräften mit und ohne VERA8-Erfahrung unterschieden. Dabei zeigten sich leichte Unterschiede bei der aufgewendeten Unterrichtszeit und der Thematisierung von Aufgabenstellungen an sich. Lehrkräfte ohne VERA8-Erfahrung sprachen weniger häufig die Informationsentnahme aus der Aufgabenstellung an. Auch bereiteten sie im Durchschnitt ungefähr eine Unterrichtsstunde weniger auf VERA8 vor und variierten minimal weniger bei den FA-Maßnahmen. Die Unterschiede wiesen dabei Effektstärken im niedrigen Bereich auf. Der Befund für Lehrkräfte ohne Erfahrung weist somit eine gewisse Ähnlichkeit zu den Befunden zu Lehrkräften auf, die als B'-Typ klassifiziert wurden. Möglich wäre nun, dass gerade diejenigen Lehrkräfte als B'-Typ klassifiziert wurden, die bisher noch keine Erfahrung mit VERA8 gesammelt haben. Tab. 7.35 gibt eine Auskunft dazu, ob die Vermutung zutrifft.

Tatsächlich beinhalten die Modellklassen von M11a_44 unterschiedlich viele Lehrkräfte ohne VERA8-Erfahrung. Die wenigsten Lehrkräfte ohne VERA8-Erfahrung finden sich in Klasse Typ G' (25.7%), gefolgt von der Klasse Typ A' (29.4%). In der Klasse Typ B' sind mit 36.0% allerdings nur leicht überdurchschnittlich viele Lehrkräfte, die über keinerlei Erfahrung mit dieser Situation verfügten. Nur die 41.6% in Klasse Typ S' sind tatsächlich deutlich überdurchschnittlich. Damit ist nicht auszuschließen, dass sich bei den beiden Gruppierungen (M11a_44 und Erfahrung mit VERA8/keine Erfahrung) Effekte überlagern.

Tabelle 7.35

eingeschätzte Veränderung der Intensität von Lehrkräften mit VERA-Erfahrung –differenziert nach Typen M11a_44

	Typ A'	Typ S'	Typ B'	Typ G'
	(n=68)	(n=154)	(n=75)	(n=74)
	absolut	Absolut	absolut	Absolut
Erfahrung	(%)	(%)	(%)	(%)
bisher keine Erfahrung mit VERA8	20 (29.4%[insg.])	64 (41.6%[insg.])	27 (36.0%[insg.])	19 (25.7%[insg.])
Intensität gleichbleibend	26 (54,1%*)	55 (61,1%*)	32 (66,7%*)	33 (60,0%*)
Intensivere Vorbereitung	9 (18,8%*)	10 (11,1%*)	6 (12,5%*)	9 (16,4%*)
Vorbereitung weniger intensiv	13 (27,1%*)	25 (27,8%*)	10 (20,8%*)	13 (23,6%*)

Anmerkung: *Die Prozentzahlen für die Zeilen „gleich bleibend", „intensiver" und „weniger intensiv" beruhen auf einem Grundwert exklusiv der Lehrkräfte ohne VERA8-Erfahrung. Die Prozentzahlen für Lehrkräfte ohne VERA8-Erfahrung beruhen auf dem Grundwert aller in dem Typ klassifizierten Lehrkräfte.

In den Klassen Typ S' und Typ B' gaben von den Lehrkräften mit VERA8-Erfahrung 61.1% bzw. 66.7% an, die Intensität ihrer Vorbereitung im Vergleich zum ersten Mal nicht verändert zu haben. In den anderen beiden Klassen waren es 60% (Typ G') und nur 54.1% (Typ A'). Der Anteil an Lehrkräften, die mit zunehmender Erfahrung glaubten, weniger intensiv vorzubereiten, ist in der Klasse Typ S' am größten (27.8%), unter den Lehrkräften des Typs B' aber am niedrigsten (20.8%).

Vergleicht man für jede der vier Klassen des Modells M11a_44 die unerfahrenen Lehrkräfte speziell mit denen, die angaben, zum Zeitpunkt der Befragung genauso intensiv auf VERA8 vorbereitet zu haben wie sie Schüler und Schülerinnen zum ersten Mal auf zentrale Lernstandserhebungen vorbereitet haben, zeigen sich für die Klassen des Typs S' (t[115]=2.161, p=033., r=.19) und des Typs B' (t[54] =2.067, p=043., r=.13) schwach signifikante Zunahmen bei der aufgewendeten Unterrichtszeit. Auch für die Klasse Typ G' erkennt man in Tab. 7.33 eine Zunahme, die aber nicht signifikant ist. Lehrkräfte des Typs A'

zeigen hingegen eine Abnahme des Stundenvolumens, die aber mit knapp neun Stunden ein sehr hohes Ausgangsniveau besitzt und auch nicht signifikant ist.[134]

Abschließend soll auch im Rahmen des Klassenmodells M11a_44 die Bedeutung von VERA8 aus Sicht der Lehrkräfte in den verschiedenen Klassen gegenüber gestellt werden. Dabei fungiert der jeweilige Wert für die Bedeutung von VERA8 für die Lehrkraft selbst als Referenzrahmen für die Bedeutung, die VERA8 nach Wahrnehmung der Lehrkräfte für die Schülerinnen und Schüler, die Schulleitung und die anderen Deutsch-, Englisch und Mathematik-Lehrkräfte besitzt. Als erstes fällt auf, dass in allen vier Klassen die Lehrkräfte annahmen, VERA8 habe für ihre Schüler und Schülerinnen eine geringere Bedeutung als für die Lehrkräfte selbst. Allerdings ist der Unterschied nicht einmal in der Klasse Typ G' signifikant.

[134] Dies steht im Widerspruch zu selbst wahrgenommenen Veränderungen der Lehrkräfte. Es ist möglich, dass Lehrkräfte, die bereits länger im Schuldienst sind, VERA8 anders wahrnehmen als Lehrkräfte, die zum Zeitpunkt der Einführung von zentralen Lernstandserhebungen in Deutschland noch in der Ausbildung waren, und daher anders vorbereiten. Möglicherweise kommt es nach dem Sammeln erster Erfahrungen mit VERA8 aber auch zu einer Angleichung des zeitlichen Vorbereitungsumfangs. Da diese Studie aber keine tatsächliche Längsschnittuntersuchung ist, kann über Entwicklungen natürlich keine Aussage getroffen werden.

Tabelle 7.36

Einschätzung der Bedeutung von VERA – differenziert nach Typen M11a_44

Bedeutung	Typ A'			Typ S'			Typ B'			Typ G'		
	n	M (SD)	95% CI	n	M (SD)	95% CI	n	M (SD)	95% CI	n	M (SD)	95% CI
für die Lehrkraft selbst	68	2.46 (1.61)	[2.09, 2.84]	155	2.59 (1.42)	[2.36, 2.82]	74	2.38 (1.43)	[2.05, 2.73]	74	2.51 (1.56)	[2.18, 2.86]
für ihre Schülerinnen und Schüler	66	2.36 (1.40)	[2.03, 2.68]	153	2.48 (1.18)	[2.28, 2.66]	75	2.36 (1.32)	[2.07, 2.64]	71	2.28 (1.31)	[1.97, 2.61]
für die Schulleitung	60	3.95 (1.42)	[3.58, 4.30]	136	3.60 (1.64)	[3.40, 3.78]	69	3.71 (1.21)	[3.41, 3.97]	66	3.68 (1.29)	[3.36, 3.97]
für andere Deutsch-, Englisch- oder Mathematiklehrkräfte der Jahrgangsstufe	62	2.47 (1.46)	[2.13, 2.84]	146	2.39 (1.12)	[2.21, 2.59]	70	2.64 (1.09)	[2.39, 2.90]	71	2.34 (1.32)	[2.00, 2.65]

Genauso schätzten die Lehrkräfte in allen vier Klassen die Bedeutung von VERA8 für die jeweilige Schulleitung als bedeutsamer ein. Hierbei beträgt der Unterschied in Klasse Typ A' aber fast eineinhalb Punkte, während der für Lehrkräfte des Typs S' durchschnittlich nur einen Punkt beträgt. Der Unterschied ist sogar schwach signifikant mit t(101.47)=1.842, p=.04, r=.13. Die Einschätzungen in den Klassen Typ B' und Typ G' liegen jeweils zwischen diesen Werten und unterscheiden sich entsprechend nicht bedeutsam von denen der anderen Klassen. Die interessantesten Unterschiede zeigen sich für den Vergleich mit den anderen von VERA8 betroffenen Kollegen. Lehrkräfte des Typs S' und des Typs G' nahmen an, VERA8 habe für ihre Kollegen im Durchschnitt eine geringere Bedeutung gehabt als VERA8 für sie selbst besaß. In der Klasse Typ A' zeigt sich hingegen praktisch kein Unterschied, während Lehrkräfte des Typs B' sogar glaubten, VERA8 sei für ihre Kollegen wichtiger gewesen als für sie selbst. Die Unterschiede in den drei Klassen sind schwach signifikant (Typ S': t[145]=2.064, p=.02, Typ B': t[68]=1.895, p=.03, Typ G': t[70]=1.688, p=.048).

Auf Grundlage des Modells M11a_44 sind damit alle relevanten Elemente ausgewertet und dargestellt. Im letzten Abschnitt dieses Kapitels folgt nun die parallel aufgebaute Darstellung der Ergebnisse auf Grundlage des Modells M22_44.

7.4 Eine nach Expertengrad differenzierte Auswertung des Vorbereitungsverhaltens – Expertenklassen für den Umgang mit Unterrichtsfeedback

Analog zu Abschnitt 7.3 und dem Modell M11a_44 aus Studie A wird in diesem Abschnitt eine Analyse des Vorbereitungsverhaltens der für das Modell M22_44 berechneten vier Klassen (Typ a, Typ b, Typ nI und Typ nII) dargestellt. Aufgrund der deutlich geringeren Fallzahl und vor allem des kleinen Umfangs der Klasse Typ nI mit nur n=22 Lehrkräften können manche Analysen allerdings nicht seriös durchgeführt werden und es muss auf diese verzichtet werden. Die Analyse beginnt wiederum mit dem zeitlichem Umfang der Vorbereitung (7.4.1) und dem inhaltliche Umfang der Vorbereitung (7.4.2), es folgen der Gebrauch von Vorbereitungs- und Kompetenzheften (7.4.3) und das außerschulische Vorbereitungsverhalten (7.4.4). Es sei noch einmal daran erinnert, dass die Klasseneinteilung von der Konzeption her nur Lehrkräfte umfassen kann, die zum Befragungszeitpunkt bereits Erfahrungen mit der Analyse von VERA8-Ergebnissen aufwiesen. Der Exkurs beschränkt sich in Abschnitt 7.4.5 daher auf die von den befragten Lehrkräften wahrgenommene Bedeutung von VERA8.

7.4.1 Zeitlicher Umfang der Vorbereitung

Insgesamt wurden mit dem Modell M22.44 N=198 Lehrkräfte klassifiziert. Für n= 187 ist eine Angabe zum für die Vorbereitung auf VERA8 genutzten Stundenvolumen vorhanden. Der Mittelwert für diese Lehrkräfte liegt bei M=7.73 Unterrichtsstunden (SE=0.37, SD=5.11, 95% CI=[7.02, 8.45]). Dies entspricht umgerechnet wiederum zwei bis zweieinhalb Schulwochen. Beim Vergleich der durchschnittlich aufgebrachten Unterrichtszeit der vier Typen des Modells M22.44 fällt auf, dass die beiden Nutzungstypen Typ nI und Typ nII den Rahmen bilden. Während Lehrkräfte des Typs nI durchschnittlich nur M=6.32 Stunden (95% CI=[4.96, 7.64]) vorbereiteten, weisen Lehrkräfte des Typs nII mit M=8.71 Stunden (95% CI=[7.36, 10.24]) den größten Wert auf. Der Unterschied ist signifikant, besitzt aber nur eine geringe Effektstärke (Welch-Test für unabh. Stichproben: t[63.76]=2.337, p=.01, r=.20). Den zweitgrößten Wert weist hier der Typ a auf mit M=8.10 Stunden (95% CI=[6.88, 9.41]), der Unterschied zu Lehrkräften des Typs nI ist aber bereits nur noch schwach signifikant und der Effekt geringer mit r=.18, t(55.52)=1.842, p=.04. Lehrkräfte des Typs b bereiteten durchschnittlich M=6.61 Stunden (95% CI=[5.36, 8.04]) vor. Die Unterschiede zu den Lehrkräften des Typs a (t[96.04]=1.526, p=.07, r=.15) und des Typs nII (t[103.68]=2.036, p=.02, r=.19) haben allerdings eine ebenfalls geringe Effektstärke.

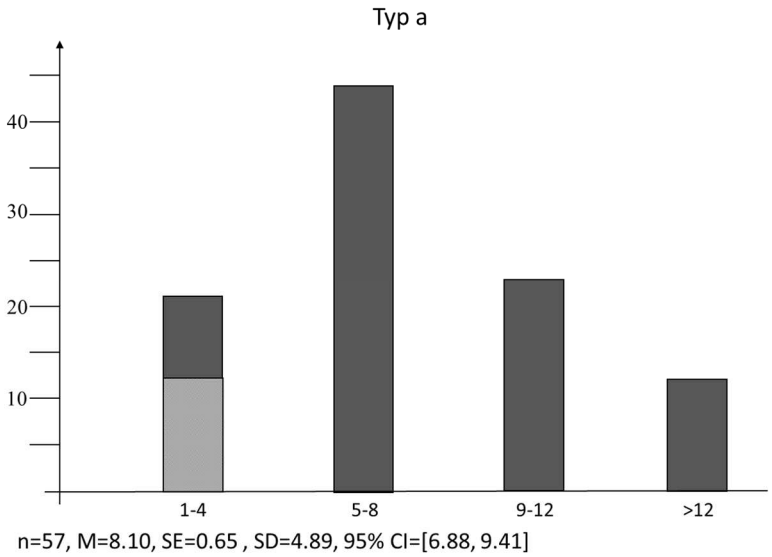
Tabelle 7.37

Anzahl der für die Vorbereitung aufgewendeten Unterrichtsstunden – differenziert nach Typen M22_44

Klasse	n	M (SD)	95% CI
a	57	8.10 (4.89)	[6.88, 9.41]
b	46	6.61 (4.94)	[5.36, 8.04]
nI	22	6.32 (3.36)	[4.96, 7.64]
nII	62	8.71 (5.76)	[7.36, 10.24]

In den nachfolgenden Abbildungen 7.23 und 7.24 ist die aufgewendete Unterrichtszeit für die vier Typen des Modells M22.44 wiederum in Form eines Histogramms dargestellt, um die Unterschiede im aufgewendeten Umfang zu verdeutlichen. Bei der Abb. 7.24 ist allerdings zu berücksichtigen, dass für diese Klassifikation wegen der geringen Fallzahl in der Klasse Typ nI eine genaue Betrachtung nach Schulwochen nur eingeschränkt aussagekräftig ist. Da die

Hypothesen H3.3 bzw. H3.4 auf eine Unterscheidung zwischen Lehrkräften abzielt, die VERA8 als Feedbackinstrument annehmen und gleichzeitig Verbesserungspotenzial haben (Typ nII), und Lehrkräften, auf die mindestens eine der beiden Bedingungen nicht zutrifft, ist es sinnvoll, die Klasse nII den drei anderen Klassen gegenüberzustellen. Es zeigt sich, dass die Vorbereitungszeit der zusammengefassten Gruppe mit M=7.24 (SE=.42, SD=4.71, 95% CI[6.44, 8.08]) über eine Stunde weniger beträgt als die der Lehrkräfte des Typs nII. Trotzdem ist der Effekt des Unterschieds mit r=.14 klein (Welch-Test für unabh. Stichproben t[102.73]=1.746, p=.04).



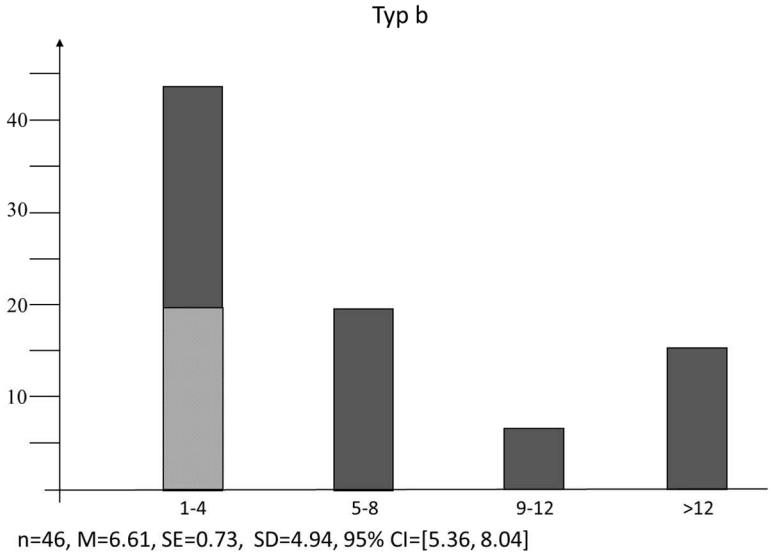
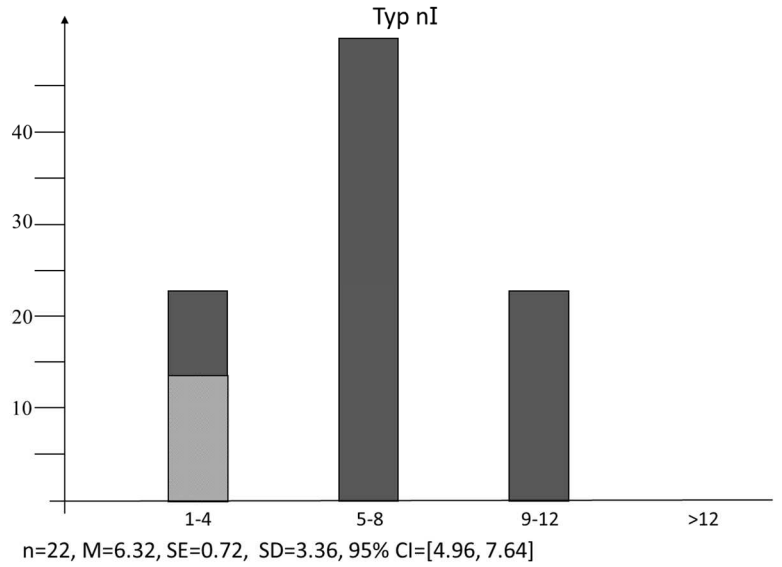


Abbildung 7.23: aufgewendetes Stundenvolumen nach Schulwochen zusammengefasst – Gegenüberstellung der Typen a und b (Verteilung der Lehrkräfte in Prozent – heller Anteil Lehrkräfte mit max. zweistündiger Vorbereitung)



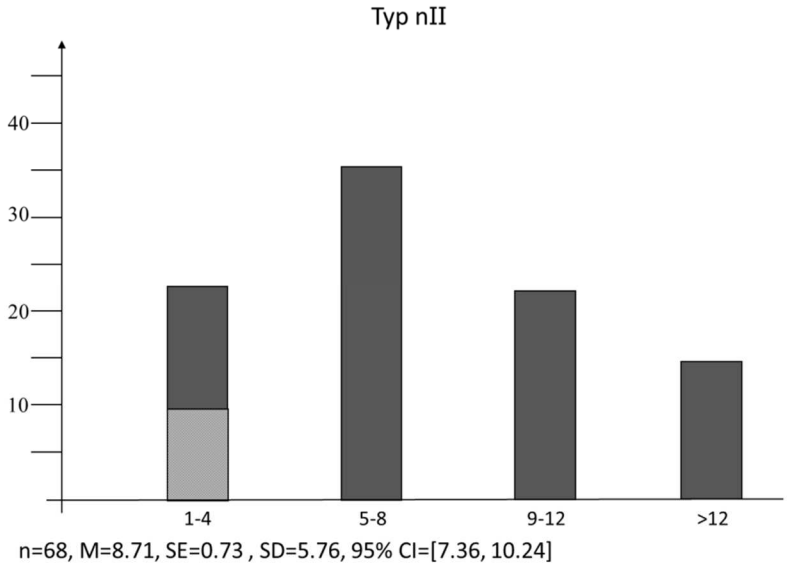


Abbildung 7.24: aufgewendetes Stundenvolumen nach Schulwochen zusammengefasst – Gegenüberstellung der Typen nI und nII (Verteilung der Lehrkräfte in Prozent – heller Anteil Lehrkräfte mit max. zweistündiger Vorbereitung)

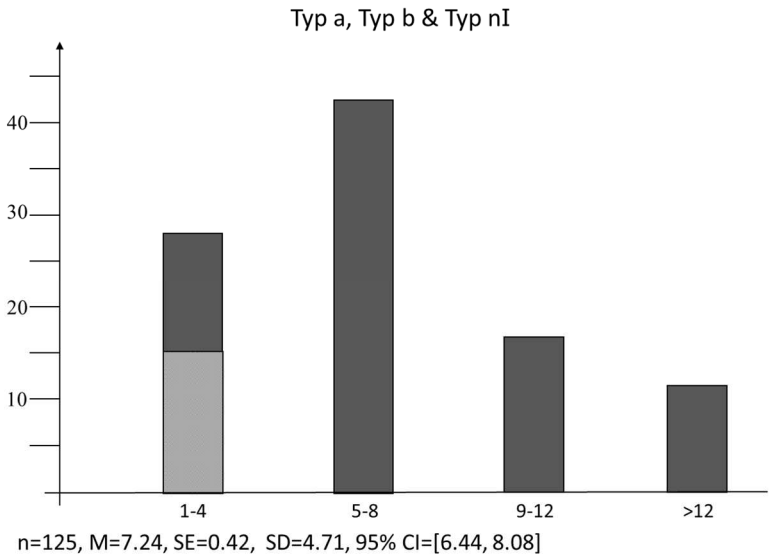
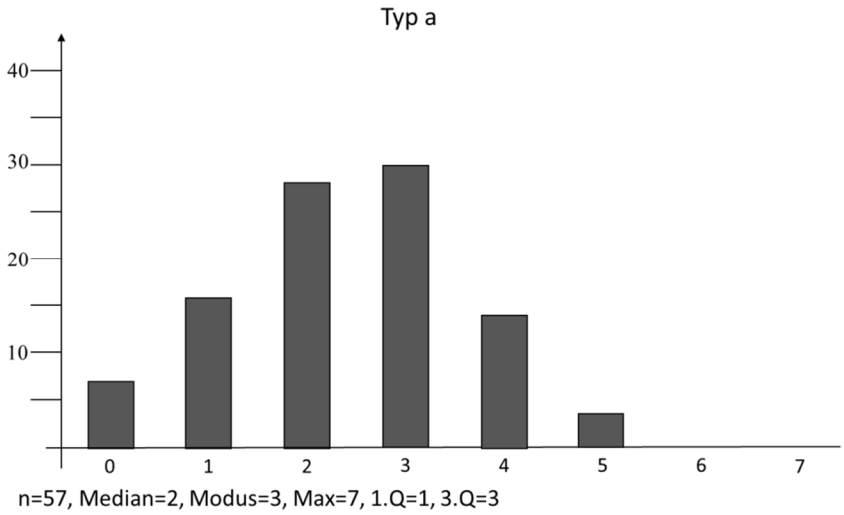


Abbildung 7.25: aufgewendetes Stundenvolumen nach Schulwochen zusammengefasst – Darstellung von Typ a, Typ b und Typ nI zusammengefasst (Verteilung der Lehrkräfte in Prozent – heller Anteil Lehrkräfte mit max. zweistündiger Vorbereitung)

7.4.2 Gestaltung der Vorbereitungszeit: Familiarity Approach, Content Approach und Test Wiseness-Strategien

Für das Modell 22_44 stellt sich das Bild der Vorbereitungsgestaltung wie folgt dar:

Von den acht Maßnahmen, mit denen Familiarity Approach (FA) durchgeführt werden kann, wurden in allen vier Klassen von den Lehrkräften mehrheitlich nur zwei oder drei Maßnahmen genutzt. Der Anteil der Lehrkräfte, die vier oder mehr Maßnahmen nutzen, beträgt zwischen 13.6% (Typ nI) bzw. 14.2% (Typ b) und 25.0% (Typ nII). Eine Lehrkraft des Typs a (hier nutzten 19.3% mehr als drei Maßnahmen) führte sieben Maßnahmen durch. Wiederum zeigt sich allein über die Kennzahlen der Maßnahmen kein deutlicher Unterschied einer Klasse zu den drei anderen, aber eine ähnliche Tendenz wie bereits bei den aufgewendeten Unterrichtsstunden. In Klasse Typ nII wurden die meisten Maßnahmen durchgeführt, als einzige weist diese Klasse bei den FA-Maßnahmen einen Median von drei auf, während er in den drei anderen Klassen jeweils nur bei zwei liegt. In der Klasse Typ nI haben Lehrkräfte angegeben, die wenigsten Maßnahmen durchgeführt zu haben.



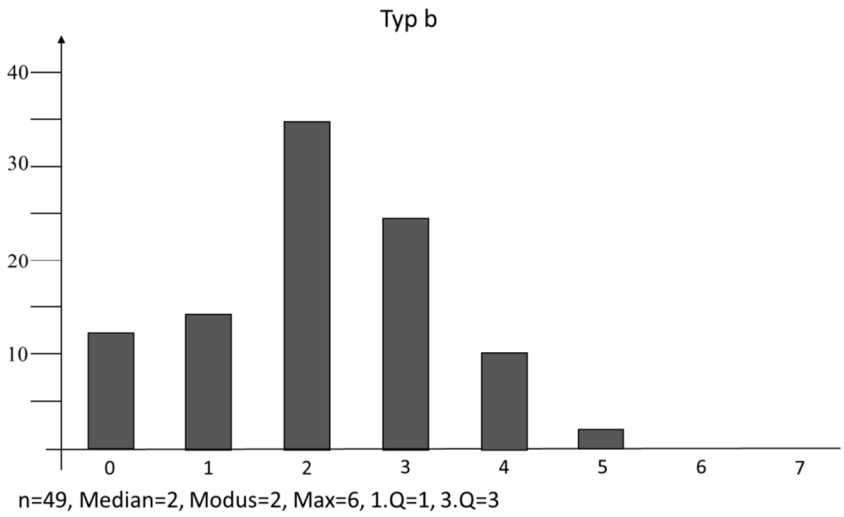
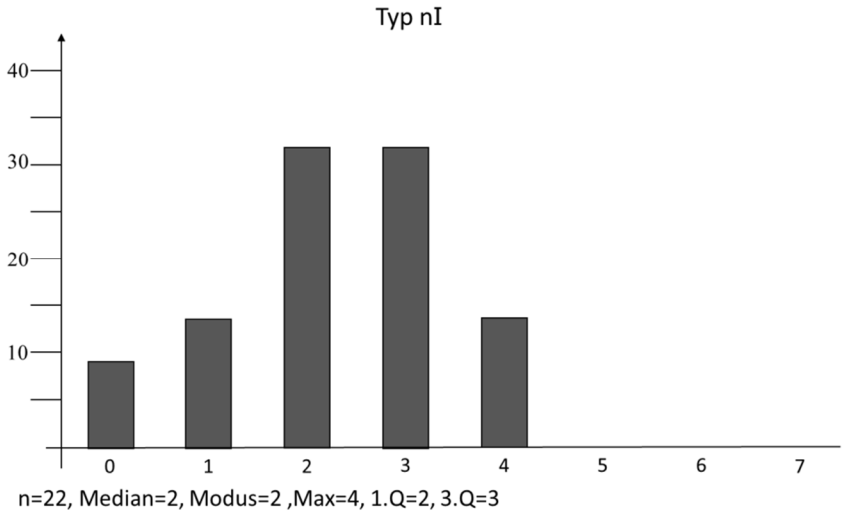


Abbildung 7.26: Darstellung der Nutzungsvariabilität von FA-Maßnahmen - Typen a und b (Verteilung der Lehrkräfte in Prozent)



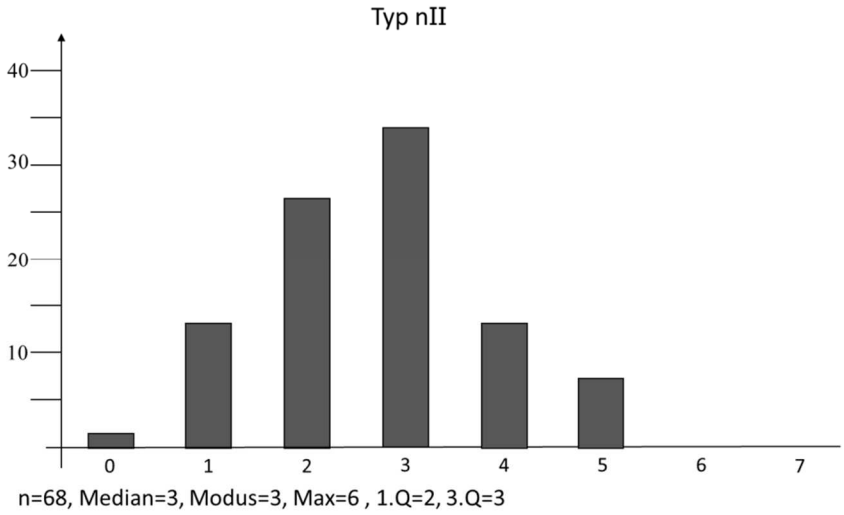


Abbildung 7.27: Darstellung der Nutzungsvariabilität von FA-Maßnahmen - Typen nI und nII (Verteilung der Lehrkräfte in Prozent)

Betrachtet man auch hier die einzelnen FA-Maßnahmen einzeln, werden Unterschiede zwischen den vier Klassen deutlich, obwohl in allen vier Klassen nur vereinzelt Lehrkräfte gar keine FA-Maßnahmen durchführten. Von den fünf abgefragten FA-Maßnahmen, die mittels Aufgaben durchgeführt werden können, ist auch in Studie B die Maßnahme (a), mit früheren VERA8-Aufgaben üben lassen, in allen vier Klassen am häufigsten von den Lehrkräften durchgeführt worden ($\chi^2[1]=2.947$, p=.01, w=.12). Während dies in den ersten drei Klassen 82.5% (Typ a), 83.7% (Typ b) und nur 77.3% (Typ nI) der Lehrkräfte machten, gaben es 91.2% der Lehrkräfte vom Typ nII an. Insgesamt weniger und in allen vier Klassen ungefähr gleich

häufig haben die Lehrkräfte zusätzlich oder alternativ solche Aufgaben zur individuellen Vorbereitung zur Verfügung gestellt. Dies waren jeweils etwas über 60%. Deutlich seltener wurden VERA8-ähnliche Aufgaben oder alte VERA8-Aufgaben in vorherige Klassenarbeiten eingebaut. Bei den VERA8-ähnlichen Aufgaben setzen sich Lehrkräfte des Typs nII aber erneut von den anderen Lehrkräften nach oben ab. Aus dieser Klasse hat immerhin jede zweite Lehrkraft ihre Klassenarbeiten entsprechend gestaltet. In den anderen Klassen sind es nur 31.6% (Typ a), 20.4% (Typ b) und 9.1% (Typ nI) der Lehrkräfte und der Unterschied ist, wenn auch nur mit mittlerem Effekt w=.28, stark signifikant ($\chi^2[1]=15.744$, p=.00). Originale Aufgaben bauten die Lehrkräfte deutlich seltener in ihre Klassenarbeiten ein (in allen Klassen unter 20%). Auch wurde nur von wenigen Lehrkräften auf Aufgaben von den offiziellen Internetseiten zurückgegriffen.

Entsprechend des allgemeinen Bildes wurden die offiziellen Internetseiten innerhalb des Unterrichts sehr selten besucht und auch eine Simulation der Testsituation haben nur vernachlässigbar wenige Lehrkräfte durchgeführt. Ihren Schülern und Schülerinnen den Besuch der offiziellen Internetseiten empfohlen haben vom bisherigen Bild abweichend besonders häufig Lehrkräfte des Typs nI. Im Vergleich zu den drei anderen Typen ist der Unterschied aber statistisch unbedeutend ($\chi^2[1]=1.335$, p=.25, w=.08).

Tabelle 7.38

Welche der folgenden Optionen haben Sie in dieser Klasse als unmittelbare Vorbereitung auf die LSE genutzt? – Familiarity Approach-Maßnahmen differenziert nach Typen M22_44

	Typ a (n=57)	Typ b (n=49)	Typ nI (n=22)	Typ nII (n=68)
Maßnahme	Absolut %	absolut %	absolut %	Absolut %
(a) mit Testaufgaben früherer LSE üben lassen	47 (82.5%)	41 (83.7%)	17 (77.3%)	62 (91.2%)
(b) alte Testaufgaben zur Verfügung gestellt	35 (61.4%)	32 (65.3%)	14 (63.6%)	46 (67.6%)
(c) zu LSE ähnliche Aufgaben in vorherige Klassenarbeiten eingebaut	18 (31.6%)	10 (20.4%)	2 (9.1%)	35 (51.5%)
(d) alte LSE-Aufgaben in Klassenarbeiten	7 (12.3%)	5 (10.2%)	4 (18.2%)	13 (19.1%)
(e) Beispielaufgaben von Homepage lösen lassen	19 (33.3%)	11 (22.4%)	6 (27.3%)	18 (26.5%)
Familiarity Approach mit Aufgaben	53 (93.0%)	43 (87.8%)	20 (90.9%)	67 (98.5%)

Maßnahme	Typ a (n=57) Absolut %	Typ b (n=49) absolut %	Typ nI (n=22) absolut %	Typ nII (n=68) Absolut %
(f) in der Klasse gemeinsam die offizielle Internetseite besucht	5 (8.8%)	3 (6.1%)	2 (9.1%)	7 (10.3%)
(g) Schüler auf offizielle Internetseite hingewiesen	32 (56.1%)	22 (44.9%)	14 (63.6%)	34 (50.0%)
(h) Testsituation simuliert	10 (17.5%)	6 (12.2%)	5 (22.7%)	12 (17.6%)

In Tab. 7.39 ist auch für die Klasseneinteilung nach M22_44 dargestellt, inwieweit bestimmte Bestandteile der Aufgaben in der Vorbereitung thematisiert wurden. Auffällig sind hier in der Tat die Ergebnisse der Klasse Typ b. Lehrkräfte aus dieser Klasse haben wesentlich seltener (p), (q) und auch (r) thematisiert. Fast die Hälfte dieser Lehrkräfte hat keine Zeit darauf verwendet zu besprechen, wie man den Aufgabenstellungen der LSE-Aufgaben die wichtigen Informationen entnimmt. Bei den drei anderen Klassen sind es nur ein Fünftel (Typ a) bzw. ein Siebtel (Typ nI und Typ nII). Für die beiden anderen Themen ist der Unterschied nicht in gleicherweise ausgeprägt, trotzdem sind es jeweils mehr als doppelt so viele Lehrkräfte aus Klasse Typ b als aus den anderen drei Klassen, die dies gar nicht thematisiert haben. Die Unterschiede zwischen der Klasse Typ b und den Lehrkräften der anderen drei Klassen sind entsprechend alle hoch signifikant mit mittlerer Effektstärke ($\chi^2_p[1]=14.736$, p=.00, w=.28; $\chi^2_q[1]=16.975$, p=.00, w=.30; $\chi^2_r[1]=5.00$, p=.03, w=.16).[135] Statistisch nicht konsistent, aber in Tab. 7.40 zu erkennen ist, dass fast alle Lehrkräfte des Typ nI alle drei Themen mindestens einmal angesprochen haben.

[135] Zu beachten ist, dass an dieser Stelle anders als bei der Analyse auf Grundlage des Modells M11a_44 nur zwischen thematisiert und nicht thematisiert unterschieden wurde, weil die Frage der umgesetzten Qualität der Vorbereitung hier untergeordnet ist.

Tabelle 7.39

Welche der folgenden Themen haben Sie im Unterricht angesprochen? – differenziert nach Typen M22_44

	Typ a (n=57)			Typ b (n=49)			Typ nI (n=22)			TypnII (n=68)		
Maßnahme	mehrfach (%)	einmal (%)	gar nicht (%)	Mehrfach (%)	einmal (%)	gar nicht (%)	mehrfach (%)	einmal (%)	gar nicht (%)	mehrfach (%)	einmal (%)	gar nicht (%)
(p) wie man die Aufgabenstellung der LSE-Aufgaben richtig versteht	40 (74.1%)	6 (11.1%	8 (14.8%)	19 (38.8%)	12 (24.5%)	18 (20.7%)	15 (71.4%)	5 (23.8%)	1 (4.8%)	43 (65.2%)	15 (22.7%)	8 (12.1%)
(q) wie man den Aufgabenstellungen der LSE-Aufgaben die wichtigen Informationen entnimmt	37 (67.3%)	7 (12.7%)	11 (20.0%)	20 (41.7%)	6 (12.5%)	22 (45.8%)	13 (61.9%)	5 (23.8%)	3 (14.3%)	38 (59.4%)	17 (26.6%)	9 (14.1%)
(r) wie man in Hinblick auf die speziellen Antwortformate richtig antwortet	38 (66.7%)	12 (21.1%)	7 (12.3%)	25 (52.1%)	10 (20.8%)	13 (27.1%)	16 (76.2%)	3 (14.3%)	2 (9.5%)	42 (63.6%)	21 (21.2%)	10 (15.2%)

Anmerkung: Die oben zu den Typen gegebene Anzahl bezieht sich immer auf die Anzahl der Lehrkräfte, zu denen mindestens zu einer der Optionen eine Angabe vorlag. Es können sich daher für die einzelnen Optionen Abweichungen in der Summe ergeben.

Die Sonderstellung von Lehrkräften des Typs b setzt sich auf den ersten Blick auch in den Ergebnissen aus der Analyse der (vorgegebenen) Maßnahmen fort, die unter Content Approach (CA) fallen. Auch hier wurde zwischen Inhalts- und Prozesskompetenzen unterschieden und es wurde abgefragt, ob diese jeweils im Unterricht angesprochen wurden und bzw. oder den Schülerinnen und Schülern als Übungsbereiche empfohlen wurden. In Tab. 7.40 sind diese Ergebnisse dargestellt. Zu beachten ist, dass wie in Tab. 7.27 unter (j*) und (m*) nur diejenigen Lehrkräfte angegeben sind, die *ausschließlich* ihrer Klasse empfohlen haben, die Inhalts- und Prozesskompetenzen zu wiederholen, dies aber nicht selbst im Unterricht gemacht haben.

Mit Ausnahme von (i) und (m) weist die Klasse Typ b in allen Zeilen jeweils den geringsten Anteil an Lehrkräften auf, die angaben, eine Vorbereitung im Sinne eines Content Approach durchgeführt zu haben. Für die Inhaltskompetenzen ist der Unterschied über alle hierzu zugeordneten Items aber statistisch nicht bedeutsam ($\chi^2_{i,j}[1]=0.816$, p=.37 w=.06). Anders sieht es für die Prozesskompetenzen aus. Diese waren in den drei anderen Klassen bei mindestens jeder zweiten Lehrkraft ein Teil der Vorbereitung (Typ a: 52.6%, Typ nI: 63.6% und Typ nII: 55.9%). Unter den Lehrkräften vom Typ b spielten die Prozesskompetenzen hingegen nur bei gut einem Viertel eine Rolle. Entsprechend ist hier der Unterschied signifikant und weist zumindest einen kleinen bis mittleren Effekt auf ($\chi^2_{k,l,m}[1]=12.590$, p=.00 w=.25). Ein relevanter Unterschied in der Frage, ob die Wiederholung von Inhalts- oder Prozesskompetenzen eher selbst in der Schule durchgeführt wurde oder den Schülern und Schülerinnen nur empfohlen wurde, lässt sich zwischen den vier Modellklassen hingegen nicht ausmachen. In allen vier Modellklassen haben mindestens so viele Lehrkräfte die Inhalts- und Prozesskompetenzen auch selbst im Unterricht wiederholt wie Lehrkräfte die Wiederholung ausschließlich in die Hände der Schülerinnen und Schüler gelegt haben.

Tabelle 7.40

Welche der folgenden Optionen haben Sie in dieser Klasse als unmittelbare Vorbereitung auf die LSE genutzt? – CA-Maßnahmen differenziert nach Typen M22_44

	Typ a (n=57)	Typ b (n=49)	Typ nI (n=22)	Typ nII (n=68)
Maßnahme	absolut (%)	absolut (%)	absolut (%)	Absolut (%)
(i) *alle* Inhaltsbereiche wiederholt	25 (43.9%)	20 (40.8%)	8 (36.4%)	30 (44.1%)
(j)* *nur* Empfohlen, alle Inhaltsbereiche zu wiederholen	16 (28.1%)	12 (24.8%)	8 (36.4%)	19 (27.9%)
alle Inhaltsbereiche wiederholt und zus. Empfohlen, alle zu wiederholen	18 (31.6%)	13 (26.5%)	7 (31.8%)	21 (30.9%)

Maßnahme	Typ a (n=57) absolut (%)	Typ b (n=49) absolut (%)	Typ nI (n=22) absolut (%)	Typ nII (n=68) Absolut (%)
(i) und/oder (j)	41 (71.9%)	32 (65.3%)	16 (72.7%)	49 (72.1%)
(k) *alle* Prozessbereiche wiederholt	18 (31.6%)	7 (14.3%)	8 (36.4%)	22 (32.4%)
(l)* *nur* empfohlen, alle Prozesskompetenzen zu wiederholen	9 (15.8%)	4 (8.2%)	5 (22.7%)	13 (19.1%)
alle Prozesskompetenzen wiederholt und zus. Empfohlen, alle zu wiederholen	9 (15.8%)	5 (10.2%)	5 (22.7%)	12 (17.6%)
(m) *eine* Prozesskompetenz besonders üben lassen	4 (7.0%)	4 (8.2%)	1 (4.5%)	8 (11.8%)
(k)+(l)+(m)	30 (52.6%)	13 (26.5%)	14 (63.6%)	38 (55.9%)
Lehrkräfte, die mindestens eine der vorformulierten CA-Maßnahmen umsetzten	46 (80.7%)	33 (67.3%)	17 (77.3%)	55 (80.9%)
Lehrkräfte, die mindestens eine der vorformulierten oder der freiformulierten CA-Maßnahmen umsetzten	47 (82.5%)	37 (75.5%)	18 (81.8%)	57 (83.8%)

Wie schon beim letzten Analysebereich (zu den Antwortformaten) sind die Werte für Lehrkräfte des Typs nII hier nicht höher als die der Lehrkräfte des Typs nI oder des Typs a. Parallel dazu gibt es auch bei den Test-Wiseness-Strategien keine Auffälligkeiten. Dies gilt für alle vier Klassen. Die Ergebnisse sind der Vollständigkeit wegen trotzdem in Tab. 7.41 aufgelistet.

Tabelle 7.41

Welche der folgenden Strategien haben Sie im Unterricht angesprochen? – differenziert nach Typen M22_44

	Typ a		Typ b		Typ nI		Typ nII	
	Ja	Nein	Ja	Nein	Ja	Nein	Ja	Nein
Test-Wiseness-Strategie	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
(TWS1)…sich nicht zu lang an einer Frage aufzuhalten	41 (73.2%)	15 (26.8%)	35 (77.8%)	10 (22.2%)	15 (75.0%)	5 (25.0%)	42 (64.6%)	23 (35.4%)
(TWS2)…sich mit den Antworten vertraut zu machen.	43 (76.8%)	13 (23.2%)	26 (60.5%)	17 (39.5%)	17 (85.0)	3 (15.0%)	53 (79.1%)	14 (20.9%)
(TWS3)…alle Antworten in Betracht zu ziehen, bevor man sich entscheidet.	40 (72.7%)	15 (27.3%)	32 (72.7%)	12 (27.3%)	14 (66.7%)	7 (33.3%)	44 (65.7%)	23 (34.3%)
(TWS4)…die Instruktionen und Fragen genau zu lesen.	56 (98.2%)	1 (1.8%)	43 (91.5%)	4 (8.5)	21 (100.0%)	0	65 (95.6%)	3 (4.4%)
(TWS5)…bei Multiple-Choice-Fragen zu raten, wenn man die Antwort nicht weiß.	17 (30.9%)	38 (69.1%)	14 (31.8%)	30 (68.2%)	6 (31.6%)	13 (68.4%)	18 (27.3%)	48 (72.7%)
(TWS6)…zuerst die Fragen zu beantworten, bei denen man sich sicher fühlt.	41 (73.2%)	15 (26.8%)	29 (63.0%)	17 (37.0%)	15 (78.9%)	4 (21.1%)	51 (76.1%)	16 (23.9%)
(TWS7)…sich spontane Einfälle zu notieren.	8 (14.5%)	47 (85.5%)	5 (11.6%)	38 (88.4%)	4 (18.2%)	14 (63.6%)	10 (15.6%)	54 (84.4%)
(TWS8)…auf grammatikalische Einschränkungen der möglichen Antworten zu achten.	23 (41.1%)	33 (58.9%)	11 (25.0%)	33 (75.0%)	10 (50.0%)	10 (50.0%)	19 (29.2%)	46 (70.8%)
(TWS9)…im Zweifel die erste Idee zu wählen, weil dies meist das Beste ist.	4 (7.7%)	48 (92.3%)	0	42 (100.0%)	1 (5.6%)	17 (94.4%)	4 (6.3%)	60 (93.8%)

7.4.3 Nutzung von Vorbereitungs- und Kompetenzheften

Bevor zum Abschluss dieses Kapitels die Analyse des Stellenwerts von VERA8 für die unter dem Modell M22_44 klassifizierten Lehrkräfte dargestellt wird, folgt die Analyse bezüglich der Nutzung von Vorbereitungs- und Kompetenzheften und ein Blick auf die außerunterrichtliche Vorbereitung. Auch unter diesem Modell müssen Vorbereitungshefte als Indikator einer input-orientierten Steuerung und Kompetenzhefte als Möglichkeit einer output-orientierten Steuerung gesehen werden. Die nach Typen differenzierte Wertung der Ergebnisse muss wiederum die grundsätzliche Tendenz berücksichtigen, dass insgesamt häufiger Vorbereitungshefte genutzt werden und damit eher eine Steuerung über Inhalte vorliegt.

Es zeigen sich bei der Gegenüberstellung der vier Klassen zwei Auffälligkeiten, die sich in das bisherige Analyseergebnis dieses Abschnitts einfügen: Lehrkräfte des Typs b (44.9%) nutzen beide Heftarten seltener als Lehrkräfte der drei anderen Klassen (Typ a: 64.2%, Typ nI: 68.4%, Typ nII: 69.2%). Dies passt zum grundsätzlichen Ergebnis für die Klasse Typ b, die die Lehrkräfte mit der geringsten Vorbereitung zusammenfasst. Demgegenüber stehen wieder die Lehrkräfte des Typs nII. Sie haben häufiger Vorbereitungshefte (68.2%) eingesetzt als ihre anders klassifizierten Kollegen und Kolleginnen (Typ a: 59.6%, Typ b: 44.9%, Typ nI: 50.0%). Die jeweiligen Effekte sind jedoch jeweils sehr klein und im zweiten Fall auch höchstens schwach signifikant mit $\chi^2(1)=4.725$, $p=.03$, $w=.16$ bzw. mit $\chi^2(1)=3.410$, $p=.07$, $w=.13$. Eine genauere Analyse der Nutzung von Vorbereitungsheften bietet sich aufgrund der geringen Fallzahlen in den Klassen Typ b und vor allem Typ nI nicht an.

Tabelle 7.42

Einsatz von Vorbereitungs- und Kompetenzheften im Unterricht - differenziert nach Typen M22_44

	Typ a (n=57)	Typ b (n=49)	Typ nI (n=22)	Typ nII (n=68)
Heftnutzung	absolut (%)	absolut (%)	absolut (%)	Absolut (%)
Vorbereitungshefte eingesetzt	34 (59.6%)	22 (44.9%)	11 (50.0%)	45 (68.2%)
Kompetenzhefte eingesetzt	12 (22.6%)	5 (11.1%)	7 (36.8%)	18 (27.7%)

Heftnutzung	Typ a (n=57) absolut (%)	Typ b (n=49) absolut (%)	Typ nI (n=22) absolut (%)	Typ nII (n=68) Absolut (%)
beide Arten eingesetzt	10 (17.5%)	4 (8.2%)	5 (22.7%)	16 (23.5%)
Lehrkräfte, die mindestens eine Heftart eingesetzt haben	34 (64.2%)	22 (44.9%)	13 (68.4%)	45 (69.2%)

7.4.4 Außerunterrichtliche Vorbereitung

Die außerunterrichtliche Vorbereitung kann als Ergänzung und als Kompensation fungieren. Da das Modell M22_44 nur eingeschränkt Ressourcen berücksichtigt, die allgemein auf die Unterrichtsgestaltung Einfluss zu haben scheinen, ergibt es sich, die außerschulische Vorbereitung als Ergänzung zu betrachten. Die Frage, ob die Schülerinnen und Schüler nach Wahrnehmung der Lehrkraft tatsächlich zu Hause besonders geübt haben, dient eher als Kontrolle, inwieweit die Lehrkraft das Verhalten der Schüler und Schülerinnen in Bezug auf VERA8 beeinflussen kann.

In Tab. 7.43 ist zu erkennen, dass Lehrkräfte des Typs b seltener empfohlen haben, sich auf VERA8 speziell vorzubereiten. Der zugehörige Effekt ist aber minimal mit w=.13, und mit $\chi^2[1]=3.470$, p=.06 nicht signifikant. Absolut gesehen ist der Wert mit 57.1% noch recht hoch. Lehrkräfte des Typs nII haben dies häufiger empfohlen (70.6%), die Werte für die beiden anderen Klassen liegen ähnlich hoch (Typ a: 70.2%) oder sogar noch darüber (Typ nI: 78.9%).

In der tatsächlichen Umsetzung der Empfehlung nähern sich die Werte für alle vier Klassen an. Mit 61.5% liegt der Wert für die Lehrkräfte des Typs b sogar als einziger über dem Wert für die Empfehlung. Leicht niedriger sind die Werte für die Klassen Typ a (60.8%) und Typ nII (60.0%). Der Wert für Lehrkräfte des Typs nI ist in diesem Fall der größte (71.4%), wobei der Unterschied erneut statistisch nicht relevant ist. Lehrkräfte dieser Analysegruppe hatten folglich den Eindruck, dass ihre Schülerinnen und Schüler sich mehrheitlich nicht an ihre Empfehlung gehalten haben.

Tabelle 7.43

außerunterrichtliche Übungsphasen in Wahrnehmung der Lehrkräfte – differenziert nach Typen M22_44

	Typ a		Typ b		Typ nI		Typ nII	
	Ja	Nein	Ja	Nein	Ja	Nein	Ja	Nein
Maßnahme	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Haben Sie der Klasse empfohlen, sich speziell auf VERA8 vorzubereiten?	40 (70.2%)	17 (29.8%)	28 (57.1%)	21 (42.9%)	15 (78.9%)	4 (21.1%)	48 (70.6%)	20 (29.4%)
Haben die Schüler für VERA8 außerhalb des Unterrichts besonders geübt?	31 (60.8%)	20 (39.2%)	24 (61.5%)	15 (38.5%)	15 (71.4%)	6 (28.6%)	39 (60.0%)	26 (40.0%)

7.4.5 Exkurs: Bewertung von VERA8

Der letzte Analyseschritt im Rahmen des Modells m22_44 befasst sich wie für das Modell M11a_44 mit der wahrgenommenen Bedeutung von VERA8 für die Schülerinnen und Schüler, die Schulleitung, ebenfalls betroffene Kollegen und Kolleginnen und für die Lehrperson selbst.

Wie auch im Exkurs zu M11a_44 ist auch in Tab. 7.44 zu erkennen, dass von allen Lehrkräften angenommen wurde, VERA8 besitze vor allem für die Schulleitung eine große Bedeutung. Interessanter ist jeweils der Vergleich des Stellenwerts von VERA8 für die Lehrkraft selbst mit den beiden anderen Vergleichsgruppen. Lehrkräfte des Typs a und auch des Typs b nahmen an, dass sie VERA8 im Vergleich zu ihren Schülerinnen und Schülern einen geringeren Stellenwert beigemessen haben. Der Wert in der Klasse Typ b ist dabei noch einmal kleiner (Lehrkräfte selbst: M=2.33, ihre Schülerinnen und Schüler: M=2.96) als in der Klasse Typ a (Lehrkräfte selbst: M=3.04, ihre Schülerinnen und Schüler: M=3.25) und auch nur im ersten Fall besteht ein signifikanter Unterschied mit t(48)=2.637, p=.01 und mittlerem Effekt r=.26. Lehrkräfte des Typs nI schätzten die Bedeutung für sich selbst und ihre Schüler und Schülerinnen durchschnittlich gleich ein (M=3.76), in der Klasse Typ nII wurde VERA8 von den Lehrkräften selbst die größere Bedeutung zugeschrieben. Der Unterschied zu ihren Schülern und Schülerinnen ist (schwach) signifikant mit t(67)=1.791, p=.04 und stellt mit r=.15 nur einen kleinen Effekt dar.

Beim Vergleich mit den ebenfalls betroffenen Kollegen gaben Lehrkräfte beider Nutzungstypen durchschnittlich an, fast keinen Unterschied zu sich selbst erkannt zu haben (Typ nI: M=3.80, Typ nII: M=3.98). Im Gegensatz dazu nahmen Lehrkräfte der anderen beiden Typen an, ihre Kollegen hätten VERA8 mehr Bedeutung beigemessen (Typ a: M=3.34, Typ b: M=2.57). In diesem Fall ist der Unterschied nur für die Klasse Typ a schwach signifikant mit t(49)=1.878, p=.04, weist aber auch hier mit r=.18 nur einen geringen Effekt auf. Beides zusammen kann als Indiz dafür genommen werden, dass die Zuordnung zu den vier Klassen gelungen ist. Auch wenn die Effekte jeweils nur einmal vorhanden und insgesamt gering sind, lässt sich doch in den unterschiedlichen Bewertungen und angenommenen Bewertungen eine erwartbare Tendenz erkennen. Es erscheint logisch, dass jemand, der VERA8 eher geringe Bedeutung beimisst, auch die Ergebnisse aus VERA8 nur bedingt einer Analyse unterzieht. Genau dies trifft nach der Konzeption auf Lehrkräfte des Typs b in jedem Fall und mit Abstrichen auch auf Kolleginnen und Kollegen des Typs a zu.

Die Ergebnisse dieser Arbeit liegen damit vollständig vor, soweit sie für die in Kapitel 5 formulierten Forschungsfragen und Hypothesen relevant sind. Damit kann im nächsten Kapitel die integrierende Diskussion dieser Ergebnisse vorgenommen und die Befunde können zusammengefasst werden.

Tabelle 7.44

Einschätzung der Bedeutung von VERA8 – differenziert nach Typen M22_44

Bedeutung	Typ a			Typ b			Typ nI			Typ nII		
	n	M (SD)	95% CI	n	M (SD)	95% CI	n	M (SD)	95% CI	n	M (SD)	95% CI
für die Lehrkraft selbst	57	3.04 (1.16)	[2.74, 3.33]	49	2.33 (1.38)	[1.96, 2.73]	21	3.76 (1.55)	[3.05, 4.38]	68	4.06 (1.38)	[3.72, 4.35]
für ihre Schülerinnen und Schüler	56	3.25 (1.10)	[2.98, 3.54]	49	2.96 (1.31)	[2.59, 3.33]	21	3.76 (1.26)	[3.19, 4.29]	68	3.81 (1.23)	[3.53, 4.10]
für die Schulleitung	52	4.58 (1.30)	[4.31, 4.83]	44	4.80 (1.30)	[4.41, 5.16]	17	5.06 (1.30)	[4.41, 5.59]	62	4.85 (1.04)	[4.60, 5.08]
für andere Deutsch-, Englisch- oder Mathematiklehrkräfte der Jahrgangsstufe	50	3.34 (1.12)	[3.04, 3.62]	44	2.57 (1.27)	[2.23, 3.62]	20	3.80 (1.06)	[2.39, 2.90]	64	3.98 (1.18)	[3.69, 4.27]

8 Diskussion, Zusammenfassung und Ausblick

Die im Jahr 2004 in Nordrhein-Westfalen unter dem Titel Lernstand9 eingeführten und ab dem Jahr 2009 im bundesweiten Rahmen als VERA8 fortgeführten jährlichen zentralen Vergleichsarbeiten sind ein Resultat der seit der Jahrtausendwende wahrgenommenen Bildungskrise als Schulleistungskrise. Wie viele andere im Zuge der Neuen Steuerung von Schule eingeführten Steuerungsinstrumente zielt auch dieses Steuerungsinstrument auf Qualitätssicherung und -entwicklung [136] ab. Auch im Zusammenhang mit zentralen Vergleichsarbeiten meint dies den Wandel zur output-orientierten Steuerung. Im Hintergrund wird folglich ein Input-Prozess-Output-Modell angenommen (Berkemeyer, 2010).

Nimmt man bisherige Forschungsarbeiten als Indikator, folgt die Einführung zentraler Vergleichsarbeiten zwei zentralen Ideen, die sich in mehreren Funktionen konkretisieren: Sie wurden *erstens* konzipiert, um den Unterricht langfristig an den in den Bildungsstandards bzw. Kernlehrplänen formulierten Zielen auszurichten, und sind somit Teil der output-orientierten Steuerung. Konkret steht hier die implementierende Funktion im Mittelpunkt, über die Testaufgaben neue Inhalte und eine neue Aufgabenkultur zu implementieren. Die Aufgaben aus den zentralen Vergleichsarbeiten sollen einerseits Indikator für das Erreichen der von administrativer Ebene intendierten Standards sein, indem sie widerspiegeln, wie weit notwendige Kompetenzen von den Schülerinnen und Schülern jeweils beherrscht werden. Andererseits sollen die Test- und Beispielaufgaben als Vorbilder für Aufgaben anderer Leistungsmessungen fungieren (erste innovierende Funktion). Die Testaufgaben müssen dadurch in irgendeiner Art Einzug in den Unterricht erhalten. Die Arbeit versuchte folglich u.a. zu beantworten, ob über zentrale Vergleichsarbeiten, also eine Form der Leistungsmessung, Unterrichtsinhalte und Aufgabentypen in den Unterricht transportiert werden können.

Zweitens wurden zentrale Vergleichsarbeiten eingeführt, um eine Evaluations- und Diagnosekultur zu initiieren (zweite innovierende Funktion). Dieser Idee liegt gleichzeitig ein Angebot-Nutzen-Modell zugrunde. Für den auf die Veränderung von Unterricht ausgelegten Evaluationszyklus sollen die angebotenen Ergebnisse aus den zentralen Vergleichsarbeiten die Ausgangsbasis sein. Im angenommenen Idealfall werden die Ergebnisse von den betroffenen Lehrkräften wahrgenommen, reflektiert und führen zu Veränderungen des Unterrichts (der dann erneut evaluiert werden kann). VERA8 ist dadurch auch ein Werkzeug zur Prozesssteuerung. Allerdings kommen bisherige Studien zu einem weniger positiven Urteil über die tatsächliche Umsetzung dieser zweiten Idee (Groß Ophoff, 2013; Hosenfeld,

[136] In diesem Kontext steht auch die Möglichkeit der Rechenschaftslegung von Lehrkräften und Einzelschulen gegenüber administrativer Ebene, Schüler und Schülerinnen sowie Erziehungsberechtigten. Der Bereich der Rechenschaftslegung findet aber in Forschungsarbeiten zu zentralen Vergleichsarbeiten in Deutschland bisher quasi keine gleichwertige Berücksichtigung.

2010; Koch, 2011; Schneewind, 2007a). Die Arbeit versuchte hieran anschließend auch einen Beitrag auf die Frage zu leisten, inwieweit dieses Angebot von Lehrkräften genutzt wird.

Beide Ideen wurden in dieser Arbeit mit einer Vorbereitung auf VERA8 verknüpft. Testcoaching im vorgelagerten Unterricht der zentralen Vergleichsarbeiten stellte den eigentlichen Schwerpunkt dar. Dabei ist eine Vorbereitung auf VERA8 im Sinne eines umfangreichen Testcoachings gerade nicht vorgesehen. Lediglich das Vertrautmachen mit den Aufgabentypen und der Testsituation kann aus der Anlage von VERA8 heraus sinnvoll sein und mögliche Übungs- und Wiederholungsphasen können insoweit angemessen sein, wie sie unabhängig von VERA8 lernförderlich sind. Mag aber eine umfangreiche Vorbereitung nicht intendiert sein, ist bzgl. der Realisierung jener Einschränkung allerdings zumindest Skepsis angebracht (Bonsen & von der Gathen, 2004, Hahn, 2008). Wenn die vorliegende Arbeit nun die Steuerungswirkung von zentralen Vergleichsarbeiten auf den vorgelagerten Unterricht (genauer: die Steuerungswirkung von VERA8 auf den vorgelagerten Mathematik-Unterricht) untersucht hat, ging es daher vor allem um nicht intendierte Effekte. Zu fragen war damit, ob gerade diese Effekte die größte Wirkung der zentralen Vergleichsarbeiten zeigen.

Sobald man diese Frage bejaht, muss man allerdings das zugrunde liegende Input-Prozess-Output-Modell hinterfragen. Theoretische Anhaltspunkte dazu sind in Kapitel 2 skizziert worden. Die Auseinandersetzung mit möglichen Defiziten des Modells auf Grundlage empirischer Daten wird explizit im letzten Abschnitt dieses Kapitels (8.3) noch einmal angestoßen. Dort werden ein entsprechendes Gesamtfazit gezogen und Perspektiven für die Praxis benannt. Davor findet in Abschnitt 8.2 als Ergänzung zur Methodendiskussion aus Kapitel 6 eine nachgelagerte kritische Auseinandersetzung mit den gewählten Wegen dieser Arbeit statt. Das Kapitel beginnt jedoch erst einmal mit der Diskussion der in Kapitel 7 dargestellten Befunde mit Blick auf die Forschungsfragen und Hypothesen. Der Abschnitt folgt der vorher getroffenen Einteilung in aggregierte Befunde zur Vorbereitung auf VERA8 in Mathematik im Jahr 2010 in Nordrhein-Westfalen und den Ergebnissen der beiden differentiellen Ansätze.

8.1 Diskussion der Ergebnisse vor dem Hintergrund der Forschungsfragen und Hypothesen

8.1.1 Umfang und Qualität von Testcoaching vor den zentralen Vergleichsarbeiten in Jahrgangsstufe 8

Zu F1: In welchem Umfang und in welcher Qualität bereiten Lehrkräfte ihre Schülerinnen und Schüler auf die zentralen Vergleichsarbeiten vor?

Erfahrungen aus anderen Ländern, insbesondere in den Vereinigten Staaten von Amerika mit High-Stake-Tests, und die Befunde der qualitativen Untersuchung zum Vorbereitungsverhalten auf Lernstand8 aus dem Jahr 2008 ließen einen hohen Vorbereitungsumfang vermuten (Amrein & Berliner, 2002; Au, 2007; Bonsen & von der Gathen, 2004; Hahn, 2008; Lind, 2009). Diese Vermutung wurde in den vorliegenden Erhebungen bestätigt: Fast alle der an einer der beiden Studien teilnehmenden Lehrkräfte haben ihre Schüler und Schülerinnen in großem Umfang vorbereitet. Der deutlichste Indikator dafür ist eine mittlere Vorbereitungszeit von zwei bis drei Unterrichtswochen. Neun von zehn Lehrkräften nutzten hierfür den naheliegenden Weg, das Training mit alten LSE-Aufgaben oder mit zu den Testaufgaben ähnlichen Aufgaben. Die Variabilität der durchgeführten Maßnahmen insgesamt war hingegen eher gering. Lehrkräfte nutzten mehrheitlich nur zwei bis drei verschiedene Maßnahmen. Festgestellt werden konnte darüber hinaus, dass Lehrkräfte die Vorbereitung vor allem während der Unterrichtszeit durchführten und ihren Schülerinnen und Schülern seltener eine Vorbereitung zu Hause empfohlen haben. Dies spricht eher gegen eine individualisierte Vorbereitung und könnte ein Zeichen dafür sein, dass die Vorbereitung von den Lehrkräften als zwingend betrachtet wurde und wohl weiterhin wird.

Unabhängig von der konkreten Ausgestaltung der Vorbereitung kann VERA8 daher ein Steuerungseffekt bescheinigt werden. Anders als durch die Bezeichnung „Teaching to the test“ konnotiert, bedeutet dieser Vorbereitungsumfang nicht zwingend, dass es sich um einen nachteiligen Effekt handelt. Die Vorbereitung kann durchaus sinnvoll genutzte Lernzeit sein (vgl. Kap 3) und dient möglicherweise dazu, Inhalte zu festigen oder angesammelte Defizite abzubauen. VERA8 ist ggf. nur der Auslöser. Gleichzeitig tritt eine Steuerungswirkung erst einmal aber unabhängig von der konkreten Unterrichtqualität der einzelnen Lehrkräfte auf und entspricht somit nicht der Intention, die Unterrichtsqualität nachhaltig zu verbessern. Von administrativer Seite ist eine Veränderung der Unterrichtsqualität des kompletten Unterrichts der dreieinhalb Jahre vor der VERA8-Erhebung anvisiert. Die Ausweisung von spezieller Vorbereitungszeit stellt eine nachhaltige Veränderung eher in Frage. Denn wäre der Unterricht nachhaltig verändert, ließe sich ein Vorbereitungsabschnitt nicht eindeutig ausweisen bzw. sollte den Lehrkräften in diesem großen Umfang obsolet erscheinen. Inwieweit zumindest die intendierte Richtung im Rahmen der Vorbereitungszeit erreicht wird, wird nachfolgend mit Rückgriff auf die Forschungsfragen und Hypothesen (vgl. Kap. 5) diskutiert.

Zu F1.1: Handelte es sich bei den in der Vorbereitung durchgeführten Maßnahmen eher um Maßnahmen, denen intendierte Effekte zugeschrieben werden können, oder eher um Maßnahmen, denen nicht-intendierte Effekte zugeschrieben werden?

Neben der Intention, eine Reflexions- und Evaluationskultur zu initiieren, stellt eine Veränderung der Unterrichtsqualität ein wesentliches Ziel von VERA8 dar. Der Steuerungseffekt auf den vorgelagerten Unterricht scheint dabei deutlich größer als der

Effekt im Sinne der zweiten innovierenden Funktion (vgl. hierzu beispielsweise Groß Ophoff, 2013). Die Einteilung von Kuper und Diemer (2011) des Nutzungsverhaltens der Daten aus zentralen Vergleichsarbeiten in eine Möglichkeit der zweckprogrammierenden Nutzung und in eine Möglichkeit der konditionalprogrammierenden Nutzung lässt sich auch auf das Vorbereitungsverhalten übertragen. Zweckprogrammierend ist eine Vorbereitung u.a., wenn sie das Resultat eines Reflexionsprozesses über die Ergebnisse aus vergangenen zentralen Vergleichsarbeiten ist, wobei angenommen werden kann, dass als Resultat einer solchen Reflexion eher eine Veränderung des Unterrichts insgesamt stehen sollte als eine bestimmte Zeit der Vorbereitung. Eine konditionalprogrammierende Nutzung kann sich in einer Vorbereitung ausdrücken, die schlechte Resultate durch eine Ad-Hoc-Vorbereitung vermeiden soll, und durch eine Vorbereitung, die unspezifisch auf schlechte Resultate folgt, ohne dass diese reflektiert wurden.

Der erhobene große Vorbereitungsumfang allein spricht eher für eine konditionalprogrammierende Nutzung. Da die Ergebnisse in Beziehung zu Vergleichsgruppen und arithmetischen Mittelwerten zurückgemeldet werden, erhalten die meisten Lehrkräfte bzw. die meisten Schulen Rückmeldungen, die Leistungen im akzeptablen Rahmen bescheinigen. Der Vorbereitungsumfang von durchschnittlich zwei bis drei Wochen lässt sich nicht als Reaktion auf schlechte Ergebnisse erklären.

Eher im Sinne einer nachhaltigen Veränderung des Unterrichts lässt sich die Integration von VERA8-ähnlichen Aufgaben oder alten LSE8-Aufgaben in Klassenarbeiten deuten. Wenn solche Aufgabentypen in Klassenarbeiten vorkommen, setzt dies voraus, dass die Schülerinnen und Schüler im Vorfeld in die Lage versetzt wurden, solche Aufgaben zu lösen. Inwieweit es sich dabei um eine zweckprogrammierende Nutzung handelt, kann nicht geklärt werden. Trotzdem lässt sich dieser Befund als Zeichen für eine Veränderung der Unterrichtsqualität verstehen, welches zumindest von dreißig Prozent der Lehrkräfte berichtet wurde.

Auch bei den Lehrkräften (80%), die sich mit den Aufgabentypen von VERA8 soweit auseinandergesetzt haben, dass sie diese speziell in der Vorbereitung angesprochen haben, kann eine zweckprogrammierende Nutzung angenommen werden. Diese besteht weniger in einer Reflexion von Ergebnisse aus vorherigen Erhebungen als aus eben jener Auseinandersetzung mit den Aufgabentypen. Auch diese Art der Reflexion ist ein intendierter Umgang und fällt unter die erste innovierende Funktion.

Für eine konditionalprogrammierende Nutzung sprechen auch die Angaben zu der thematischen Zielrichtung der Vorbereitungsphase. Einem große Anteil an Lehrkräften, die angegeben haben, in der Vorbereitungsphase alle Inhalts- bzw. alle Prozesskompetenzen wiederholt zu haben, steht nur ein geringer Anteil von unter 15% der Lehrkräfte gegenüber, die ihre Vorbereitung spezifisch auf die Wiederholung einzelner Prozesskompetenzen ausgerichtet haben.

Zu F1.2: Inwieweit wird eine Testkompetenz vermittelt, die die Testvalidität erhöht?

Grundsätzlich erhöhen alle Maßnahmen die Testvalidität, die unter den Bereich Familiarity Approach gefasst werden können. Je mehr sich die Schüler und Schülerinnen ausschließlich auf die Inhalte der Testaufgaben konzentrieren können und je weniger sie ihre Ressourcen für das Verständnis neuartiger Aufgabentypen einsetzen müssen, desto besser erfassen Tests wie die von VERA8 die zu prüfenden inhaltlichen und prozessbezogenen Kompetenzen. Auch einige Test Wiseness Strategien können dazu dienen, die Testvalidität zu erhöhen (siehe im Abs. 7.4 die Strategien der Gruppen 1 & 3).

Für die Maßnahmen und Thematisierungen in Bezug auf die Aufgabentypen kann ein insgesamt breites Spektrum zu einem Familiarity Approach festgestellt werden. Dabei zeigten sich erwartbare Differenzen in der Nutzungshäufigkeit der Maßnahmen: Ersten wurden einfach zugängliche Maßnahmen – allen voran eine Übungsphase mit alten Testaufgaben – und Themen deutlich häufiger genutzt als aufwendigere Maßnahmen wie die Simulation der Testsituation durchgeführt wurden. Auch wurden weniger zugängliche Bereiche der Testkompetenz wie die Informationsentnahme aus der Aufgabenstellung seltener thematisiert. Zweitens wurde die Vorbereitung eher im Klassenverband durchgeführt, soweit es sich dabei um das Üben mit Testaufgaben handelt, aber eher den Schülerinnen und Schülern für die außerschulische Vorbereitung überlassen, wenn es um die vorwiegend mit Onlinepräsenzen mögliche Informationsbeschaffung über Abläufe und Ziele ging. Die geringe Nutzungsvariabilität kann hingegen darauf deuten, dass der demgegenüber hohe Vorbereitungsumfang (gemessen in Unterrichtsstunden) weniger zielgerichtet für ein FA genutzt wurde.

Für die Test Wiseness Strategien lässt sich festhalten, dass Lehrkräfte vor allem diejenigen Strategien vermittelten bzw. angesprochen haben, die nicht nur zu besseren Leistungen der Schüler und Schülerinnen in der Bearbeitung des Tests führen sollen, sondern auch mit dem Zweck des Tests vereinbar sind. Dies sind die TWS aus den Gruppen 1 und 3 (s. 7.1). TWS aus der Gruppe 2 (s. 7.1) wurden von einem geringeren Anteil der Lehrkräfte vermittelt. Dies könnte derart gedeutet werden, dass Lehrkräfte ihren Schülerinnen und Schülern auch nicht den Eindruck vermitteln wollten, bei Tests (zumindest dieser Art) komme es nur darauf an, erfolgreich abzuschneiden, sondern VERA8 als Kompetenzmessung ernst nahmen und dies auch ihren Schülerinnen und Schülern so weitergeben wollten.

Die Testsituation als besondere Testsituation steht hingegen nur bei jeder zweiten Lehrkraft im Fokus. VERA8 scheint für Mathematik-Lehrkräfte vorwiegend eine zentral gestellte Klassenarbeit zu sein, die möglicherweise Inhalte abfragt, die vorher nicht im Unterricht (ausreichend) behandelt worden sind. Ein Testcoaching im Sinne eines Familiarity Approach wurde daher eher als zusätzlicher Effekt der durchgeführten Maßnahmen realisiert, ohne dass eine höhere Testvalidität dabei tatsächlich als Ziel angenommen werden muss. Trotzdem sind die Schüler und Schülerinnen der befragten Lehrkräfte mehrheitlich ausreichend in diese Richtung vorbereitet worden. Die Testvalidität von VERA8 ist in diesem Punkt entsprechend hoch.

Zu F.1.3: Inwieweit wird die Vorbereitung genutzt, wichtige fachliche Inhalte zu vermitteln bzw. zu wiederholen?

Drei von vier Lehrkräften gaben an, die Vorbereitungsphase genutzt zu haben, um Inhalte zu üben oder zu wiederholen. VERA8 fungierte hier folglich als Anlass für diese aus fachdidaktischer Sicht wichtige Übungs- und Wiederholungsphase. Die Ergebnisse der beiden Studien zeigen aber auch, dass vorwiegend Inhaltskompetenzen im Vordergrund stehen. Prozesskompetenzen allgemein haben weniger als dreißig Prozent der Lehrkräfte wiederholt, spezifisch einzelne Prozesskompetenzen haben nur knapp zehn Prozent der Lehrkräfte in den Fokus genommen. Dabei spielten Prozesskompetenzen gerade in der Vorbereitungsphase derjenigen Lehrkräfte eine größere Rolle, die auch gezielt oder allgemein Inhaltskompetenzen wiederholt haben bzw. diese üben ließen. Die implementierende Funktion mit Bezug zur Veränderung der Unterrichtsqualität durch neue Aufgaben von VERA8 kann daher nur eingeschränkt als verwirklicht bezeichnet werden. Zwar werden offenbar Übungs- und Wiederholungsphase initiiert, es sind aber offensichtlich Übungs- und Wiederholungsphasen zu den traditionellen Inhaltskompetenzen. Den Stellenwert von Prozesskompetenzen im Bewusstsein der Lehrkräfte scheint VERA8 (noch) nicht erhöhen zu können. Darüber hinaus spielen Übungs- und Wiederholungsphasen für zwanzig Prozent der Lehrkräfte überhaupt keine bewusste Rolle.

Zu F1.4: Inwieweit hat vorherige Erfahrung mit zentralen Vergleichsarbeiten in den Jahren davor einen Einfluss auf die Intensität der Vorbereitung und werden Schwerpunktsetzungen mit Blick auf VERA8 vorgenommen?

Der Vergleich von Lehrkräften ohne VERA8-Erfahrung mit Lehrkräften, die in anderen Jahren davor bereits an VERA8 teilgenommen haben, zeigte Unterschiede zwischen den beiden Gruppen. Selbst wenn man nur explizit diejenigen Lehrkräfte mit Erfahrung als Referenz nimmt, die von sich selbst behaupten, die Intensität der Vorbereitung nicht verändert zu haben, bereiteten VERA8-erfahrene Lehrkräfte durchschnittlich eine Unterrichtsstunde mehr vor. Ebenso gaben von jenen Lehrkräften mit VERA8-Erfahrung mehr an, Übungs- und Wiederholungsphasen in der Vorbereitungsphase durchgeführt zu haben als ihre Kolleginnen und Kollegen ohne VERA8-Erfahrung. Dabei gaben die Lehrkräfte die Verantwortung für die Übungs- und Wiederholungsphasen weniger in die Hand der Schüler, sondern führten diese tatsächlich in der Unterrichtszeit durch. Für die FA-Maßnahmen konnte diese Tendenz zwar nicht bestätigt werden, dies könnte sich allerdings auch daraus ergeben, dass für ein Testcoaching als Familiarity Approach recht schnell der maximale Effekt erreicht scheint.

Wiederum in Richtung des hier dargestellten Trends liegen die Ergebnisse zu der Thematisierung einzelner Aspekte der Aufgabenstellungen. Auch hier investierten erfahrene Lehrkräfte mehr in die drei Themen (p) wie man die Aufgabenstellung der LSE-Aufgaben richtig versteht, (q) wie man den Aufgabenstellungen der LSE-Aufgaben die wichtigen

Informationen entnimmt und (r) wie man in Hinblick auf die speziellen Antwortformate richtig antwortet. Die Unterschiede der Thematisierung einzelner Aspekte der Aufgabenstellungen lassen zwei Schlüsse zu: Erstens ist unerfahrenen Lehrkräften die grundsätzliche Relevanz der Themen für ein erfolgreiches Abschneiden der Schüler und Schülerinnen bei VERA8 (und ggf. auch bei nachfolgenden Tests) weniger bewusst und zweitens halten diejenigen unerfahrenen Lehrkräfte, die eine gewisse Relevanz erkennen, die einmalige Thematisierung für ausreichend, während erfahrene Lehrkräfte eher dazu neigen, die drei Themen mehrfach anzusprechen. Mit besonderem Fokus auf (q) kann man weiterhin sehen, dass bei dem schwierigsten aufgabenbezogenen Thema der Unterschied zwischen erfahrenen und unerfahrenen Lehrkräften im hier beschriebenen Sinne besonders deutlich wird. Noch einmal deutlich weniger unerfahrene Lehrkräfte sprachen überhaupt an, wie man die wichtigen Informationen aus der Aufgabenstellung entnimmt.

Auch insgesamt liegt es nahe, die höhere Vorbereitungsintensität der Lehrkräfte mit VERA8-Erfahrung als Resultat der Auseinandersetzung mit vorherigen Erhebungen zu erklären. Dafür spricht auch die lehrerzentrierte Engführung im Unterricht selbst bei den Übungs- und Wiederholungsphasen. Die alternative Erklärung, neuere Lehrkräfte seien besser über VERA8 informiert und ihnen sei deswegen bewusst, dass eine (intensive) Vorbereitungsphase unnötig und der Testvalidität eher abträglich sei, sodass sie deswegen weniger intensiv vorbereiteten, scheidet aus. Dazu ist auch in dieser Gruppe der durchschnittliche Vorbereitungsumfang viel zu groß.

In Bezug auf eine mögliche Schwerpunktsetzung zeigte sich nur bei einem geringen Teil der Lehrkräfte und bei einem noch geringeren Anteil der Fachgruppen eine Veränderung. Veränderungen zu einer intensiveren Behandlung einer (Prozess-)Kompetenz waren häufiger als umgekehrt ein weniger intensives Unterrichten einer (Prozess-)Kompetenz im Vergleich zu den Schuljahren davor. Nur von einem kleinen Teil der Lehrkräfte wurden diese Veränderungen auf Überlegungen zurückgeführt, die im Zusammenhang mit VERA8 standen. Die Ergebnisse der Interviewstudie aus dem Jahr 2008 konnten in den beiden Studien 2010 daher nicht repliziert werden. Die vormals beobachtete Ausrichtung auf eine bestimmte Prozesskompetenz scheint dadurch allein über den angekündigten Testschwerpunkt begründet gewesen zu sein. Schwerpunktsetzungen sind folglich vor allem Ausdruck einer konditionalprogrammierenden Nutzung von VERA8.

Zu F1.5: Inwieweit spricht der Umfang der Nutzung von Vorbereitungsheften und Kompetenzheften/Lernhilfen eher für eine output- oder eine input-orientierte Steuerung?

Die Einführung von zentralen Vergleichsarbeiten geschah im Zuge der Umstellung auf eine output-orientierte Steuerung. Statt genaueren Vorgaben für die Unterrichtsinhalte stellt die administrative bzw. die politische Ebene Vorgaben für die Unterrichtsziele bereit. Zentrale Vergleichsarbeiten drücken den Paradigmenwechsel insoweit aus, dass sie vorgeben, das Erreichen der Ziele zu überprüfen, den Weg dorthin aber den einzelnen Schulen und

Lehrkräften zu überlassen. Insbesondere die von den Schulbuchverlagen angebotenen Vorbereitungshefte, aber auch unter bestimmten Zwecken genutzten Kompetenzhefte/Lernhilfen stellen diesen Paradigmenwechsel allerdings in Frage, wenn Lehrkräfte sich durch die Einführung von zentralen Vergleichsarbeiten wie VERA8 zur Nutzung dieser Hefte genötigt sehen. Eine freie Wahl der Mittel, um die administrativ vorgegebenen Unterrichtsziele zu verwirklichen, ist dann nicht mehr gegeben, sodass von einer neuen input-orientierten Steuerung ausgegangen werden muss.

Die Ergebnisse aus dem Abschnitt 7.1.7 deuten in der Tat darauf hin, dass Lehrkräfte mehrheitlich annehmen, diese Hefte nutzen zu müssen. Drei von fünf Lehrkräften haben mindestens eine der beiden Heftarten eingesetzt. Die Mehrzahl nutzte wie erwartet Vorbereitungshefte. Da sich diese sehr stark an den konkreten Aufgabenstellungen von VERA8 orientieren, rücken die zugrunde liegenden mathematischen Kompetenzen in den Hintergrund. Die mit diesen Heften durchgeführte Vorbereitung erscheint dadurch eher test- orientiert als kompetenz-orientiert.

Gleichzeitig zeigen die Ergebnisse der beiden Studien aber auch, dass diejenigen Lehrkräfte, die für die Vorbereitungsphase auf diese Heftarten zurückgriffen, gezieltere Übungs- und Wiederholungsphasen durchführen ließen. Eher wurde mit den Heften angeregt, *einzelne* Prozess- und noch häufiger *einzelne* Inhaltskompetenzen zu wiederholen, statt allgemein alle Kompetenzen wiederholen zu lassen.

8.1.2 Testcoaching als Form der Unterrichtsqualität unter dem Lehrer-Expertenansatz

Zu F2: Lassen sich Unterschiede zwischen Lehrkräften in Umfang und Qualität der Vorbereitung durch die Typisierung der Lehrkräfte nach personenbezogenen Überzeugungen erklären?

Über die allgemeinen deskriptiven Befunde zu den Vorbereitungsphasen als Steuerungseffekte von VERA8 hinausgehend wurden in dieser Arbeit zwei differentielle Ansätze verfolgt, um Unterschiede im Vorbereitungsverhalten durch Gruppierungen der beteiligten Lehrkräfte zu erklären. Der erste differentielle Ansatz versucht eine Verbindung zwischen der Unterrichtsqualität der Vorbereitungsphase und personenbezogenen Überzeugungen herzustellen. Grundlage dieser ersten differentiellen Analyse war das Modell zur Lehrer-Handlungskompetenz von Baumert & Kunter (2006) in modifizierter Form, welches u.a. auf dem Job Demand-Rescource-Model von Schaufeli u.a. (Schaufeli & Bakker, 2004) bzw. auf Befunden von Schaarschmidt & Fischer (Schaarschmidt & Fischer, 2008) beruht, aber auch personenbezogene Ressourcen wie Kontroll- und Kompetenzüberzeugungen beinhaltet.

Die Grundannahme dieses ersten differentiellen Zugangs besteht darin, dass auch die Vorbereitungsphase von Lehrkräften als wichtiger Teil des Unterrichts gesehen wird und eine gute Vorbereitung als Anforderung (sei es von Seite der Eltern sowie Schülerinnen und Schüler, der Schulleitung oder von administrativer Seite) akzeptiert wird, die es zu bewältigen gilt. Entsprechend sollte die Bewältigung dieser Anforderung gerade den Lehrkräften am besten gelingen, die auch für den sonstigen Unterricht die besten Voraussetzungen mitbringen.

Es wurden für verschiedene Modellausschnitte latente Klassenanalysen durchgeführt. Der Vergleich der Modellparameter wies für das Klassenmodell M11a mit einer Vier-Klassen-Lösung für die sechs (Kurz-)Skalen berufliches Beanspruchungserleben [BEL], Overcommitment [OC], Arbeitsengagement [UWES], fachdidaktisches Fähigkeitsselbstkonzept [SK] und Selbstwirksamkeitserwartung [SWE] sowie Gewissenhaftigkeit [GW] die besten Werte auf. Weitere latente Klassenanalysen wurden auch für das Klassenmodell M11 berechnet, in das nur die ersten fünf der sechs (Kurz-)Skalen Eingang fanden. Die Drei-, Vier- und Fünf-Klassen-Lösungen wiesen aber keine akzeptablen Passungswerte auf.

Die berechnete Vier-Klassen-Lösung für das Klassenmodell M11a (inklusive der Skala GW) ergibt auch inhaltlich eine größere Übereinstimmung zu den Mustern des AVEM von Schaarschmidt & Fischer. Dies lässt sich u.a. dadurch begründen, dass das Arbeitsengagement einen Schwerpunkt auf die Arbeitsbereitschaft legt und daher durch die Skala GW besser abgebildet wird als allein durch die Skala UWES, die mehr auf die Arbeitszufriedenheit zielt. Entsprechend ergaben sich vier Muster (Typ A‘, Typ B‘, Typ G‘ und Typ S‘), die eine im gegebenen Rahmen mögliche Übereinstimmung mit den vier Mustern des AVEM besitzen. Auf die Muster Typs A‘, Typs B‘ und Typs G‘ entfielen jeweils ca. 20% der Lehrkräfte, auf das Muster Typs S‘ entsprechend doppelt so viele. Das Muster Typ A‘ zeichnete sich durch die höchsten Werte in der Arbeitsbereitschaft und den Kontroll- und Kompetenzüberzeugungen einerseits und durch negative berufliche Emotionen andererseits aus. Im Muster Typ B‘ sind vorwiegend Lehrkräfte klassifiziert, die generell bedrohlich schlechte Werte in allen sechs Bereichen aufwiesen. Umgekehrt zeigten Lehrkräfte im Muster Typ G‘ in allen sechs Bereichen sehr gute oder gute Werte. Das Muster Typ S‘ fasst schließlich diejenigen Lehrkräfte zusammen, die im Bereich der beruflichen Emotionen gute Werte besaßen, in den anderen Bereichen (zum Teil deutlich) aber schlechtere Werte zeigten.

Neben den Klassenmodellen M11 und M11a wurden auf Grundlage des Datensatzes aus Studie A auch latente Klassenanalysen für ein Klassenmodell M12 berechnet, das zusätzlich Kausalüberzeugungen über das Lehren & Lernen berücksichtigt. Die dort erhaltenen Klassenlösungen erwiesen sich aber als nicht interpretierbar und zeigten darüber hinaus ebenfalls keine ausreichenden Passungswerte. Inwieweit die Akzeptanz der einzelnen fachlichen Unterrichtsziele durch die Lehrkräfte einen Einfluss auf die Gestaltung der Vorbereitungsphase hat, konnte daher nicht untersucht werden. Es wurde folglich nur das

Klassenmodell M11a für weitere Analysen herangezogen. Die Befunde aus der Studie A sollen nachfolgend in Bezug zu den in Kap. 5 aufgestellten Hypothesen diskutiert werden.

Zur Hypothese H2.1 – mehr Engagement der Lehrkraft führt zu größerem Vorbereitungsumfang:

Es wurde angenommen, dass Lehrkräfte, die überdurchschnittlich engagiert sind (Typ G‘ oder Typ A‘), umfangreicher auf die Vergleichsarbeiten vorbereiten als Lehrkräfte, die über weniger Arbeitsengagement verfügen (Typ B‘ und Typ S‘) (Hypothese 2.1). Ein großes Arbeitsengagement gehört als Kombination aus Arbeitszufriedenheit und Arbeitsbereitschaft in jedem Fall zu den förderlichen Ressourcen, um den an Lehrkräfte gestellten Anforderungen gut begegnen zu können. Entsprechend konnte ein Zusammenhang von hohem Vorbereitungsumfang und hohem Arbeitsengagement erwartet werden.

Die Befunde in Studie A scheinen die Erwartungen zu bestätigen. Lehrkräfte des Musters Typ G‘ bereiteten ihre Schülerinnen und Schüler durchschnittlich mit dem größten zeitlichen Umfang (über acht Unterrichtsstunden) vor. Lehrkräfte der Muster Typ G‘ und Typ A‘ bereiteten durchschnittlich über eine Stunde mehr auf VERA8 vor als die Lehrkräfte der anderen beiden Muster (durchschnittlich knapp sieben Unterrichtsstunden). Die Effekte sind allerdings nur schwach signifikant und sehr gering. Lehrkräfte der ersten beiden Muster bereiteten häufiger (Modalwert) zwei Wochen vor, während Lehrkräfte der anderen beiden Muster häufig nur bis zu vier Unterrichtsstunden vorbereiteten. Gleichzeitig zeigte sich aber bei keinem Muster ein größerer Anteil an Lehrkräften, die nur im sinnvollen Rahmen von maximal zwei Unterrichtsstunden vorbereiten. Allein auf den Indikator der aufgewendeten Unterrichtszeit beschränkt kann daher geschlossen werden, dass überdurchschnittlich engagierte Lehrkräfte umfangreicher auf zentrale Vergleichsarbeiten vorbereiten.

Die gleiche Tendenz zeigte sich zwar auch für den Indikator der Anzahl an genutzten FA-Maßnahmen. Allerdings war der Unterschied noch geringer und besitzt keine Aussagekraft. Überhaupt keinen nennenswerten Unterschied gab es zwischen den vier Mustern bei der Empfehlung einer außerunterrichtlichen Vorbereitung.

Für das Muster Typ A‘ zeigt sich die angenommene herausgehobene Stellung folglich nicht. Zwar bereiteten Lehrkräfte auch aus diesem Muster umfangreicher vor, aber sie bereiteten nicht am umfangreichsten vor, obwohl dieses Muster gerade Lehrkräfte mit besonders großer Arbeitsbereitschaft und sehr guten Kontroll- und Kompetenzüberzeugungen umfasst. Diese allein scheinen daher noch kein ausreichender Ressourcen-Pool zu sein, um in Kombination mit einer erlebten beruflichen Unzufriedenheit eine besonders umfangreiche Vorbereitungsphase zu VERA8 zu praktizieren.

Zur Hypothese H2.2 – größere personenbezogene Ressourcen führen nicht zu mehr Vorbereitungsqualität:

Unter der Folie des Modells der Lehrer-Handlungskompetenz ist in den Kapiteln 4 & 5 herausgearbeitet worden, dass Lehrkräfte mit sehr großen Ressourcen eine qualitativ bessere Vorbereitung gestalten sollten (Hypothese 2.2). Dies sollte sich vor allem in einem Vergleich der Muster Typ G‘ und Typ B‘ zeigen.

In der Tat offenbarten sich Unterschiede zwischen den beiden Mustern in den FA-Maßnahmen: Lehrkräfte des Typs B’ nutzten auch – wenn auch weniger häufig – Aufgaben aus zentralen Lernstandserhebungen im Unterricht, der Aufgabentypus fand aber seltener Niederschlag in den anderen Teilen der Leistungsmessung. Geht man davon aus, dass die Aufgaben von schriftlichen Arbeiten und anderen Formen der Leistungsmessung ein Abbild der Unterrichtsinhalte darstellen, lassen diese Unterschiede in den Klassenarbeiten weiter auf Unterschiede im Unterricht schließen (erste Innovierende Funktion). Die Aufgaben werden folglich wohl eher zum Üben als Vorbereitung für VERA8 genutzt, scheinen aber weniger Teil des restlichen Unterrichts zu sein. Besonders Maßnahme (c), zu VERA8 ähnliche Aufgaben in vorherige Klassenarbeiten einzubauen, kann hier als Indikator gelten. Dies ist nicht mit einer Aussage über die grundsätzliche Unterrichtsqualität gleichzusetzen. Aufgaben stellen lediglich ein Mittel dar, Schülerinnen und Schülern die notwendigen Lerngelegenheiten zu geben. Gleichwohl kann in Frage gestellt werden, warum diese speziellen Aufgaben von jenen Lehrkräften des Typs B’ nur additiv genutzt wurden und nicht Teil der Leistungsmessungen waren.

Auf Grundlage von allgemeinen Merkmalen hoher Unterrichtsqualität von Brophy (1999), Helmke (2009) oder Meyer (2009) wurden vier Elemente als Merkmale einer qualitativ höheren Vorbereitungsphase herausgearbeitet:

Das erste wichtige Merkmal einer qualitativ höheren Vorbereitungsphase ist eine anspruchsvollere Vorbereitung. Hierunter fällt u.a. der Bereich der Aufgabentypen. Die Befunde zu diesem Bereich zeigen aber für Lehrkräfte des Typs G‘ keine bemerkbar häufigere Thematisierung der drei angefragten Maßnahmen. Im Gegenteil gaben diese Lehrkräfte sogar häufiger als alle anderen an, die Entnahme der wichtigsten Informationen überhaupt nicht angesprochen zu haben.

Das zweite wesentliche Merkmal einer qualitativ höheren Vorbereitungsphase drückt sich in dem Bereich der Übung und Wiederholung aus. Die qualitativ höhere Vorbereitungsphase umfasst neben den inhaltsbezogenen Kompetenzen auch prozessbezogene Kompetenzen. Dies zeigte sich für Lehrkräfte des Musters Typ G‘ in der Tat. Sie wiederholten neben allen Inhaltsbereichen auch gezielt einzelne oder alle Prozesskompetenzen oder ließen diese in der Unterrichtszeit üben. Dies gilt aber in gleicher Weise auch für Lehrkräfte des Typs A‘, sodass hier kein Alleinstellungsmerkmal des Musters Typ G‘ vorliegt.

Weiter kann sich eine qualitativ gute Vorbereitung dadurch auszeichnen, dass die Vorbereitung Schülerinnen und Schülern Freiraum in der Vorbereitung lässt und sie gleichzeitig unterstützt. In der zugrunde liegenden Studie konnte sich dieser Qualitätsunterschied vor allem darin zeigen, dass Inhalts- und Prozesskompetenzen zur Wiederholung empfohlen wurden, gleichzeitig aber auch im Unterricht thematisiert wurden. Tendenziell gilt dies zwar wiederum für die Muster Typ G‘ und mehr noch Typ A‘, die Anzahl der Lehrkräfte, die dies aber derart umgesetzt zu haben scheinen, sind aber insgesamt in der Minderzahl und die Unterschiede zu den anderen beiden Mustern nicht statistisch bedeutsam. Häufiger als von Lehrkräften der Muster haben die Schüler und Schülerinnen einen Hinweis auf die Homepages zu VERA8 von Lehrkräften des Typs G‘ erhalten. Dies bietet den Schülerinnen und Schülern zumindest die Möglichkeit, sich selbstständig weitergehend über VERA8 zu informieren und mit dem Instrument vertraut zu machen.

Schließlich kann eine variantenreiche Vorbereitungsphase als qualitativ höher bezeichnet werden. Die leicht höhere Anzahl an Lehrkräften aus den Mustern Typ G‘ & Typ A‘, die vier oder mehr verschiedene FA-Maßnahmen durchführten, könnte ein Anzeichen für mehr Variabilität sein. Aber auch hier besitzt das Muster Typ G‘ keine exklusive Stellung und die Unterschiede zu den beiden anderen Mustern sind minimal.

Für das Muster Typ G‘ kann anhand dieser vier Merkmale die Erwartung über die Qualität der Vorbereitungsphase nicht bestätigt werden. Wenn Unterschiede zu den Mustern mit besonders schlechter Ausgangslage zu finden waren, galten sie jeweils mindestens genauso auch für Lehrkräfte des Typs A‘. Es waren darüber hinaus jeweils nur kleine Unterschiede. Gleichzeitig ist aber mit der unterschiedlichen Nutzung von Testaufgaben und VERA8-ähnlichen Aufgaben in vorherigen Klassenarbeiten auch ein Hinweis darauf gegeben, dass der Steuerungseffekt von VERA8 auf Lehrkräfte des Typs G‘ positiver ist als beispielsweise auf Lehrkräfte des Typs B‘.

Zur Hypothese H2.3 – Teaching to the Test trotz positiver Kontrollüberzeugungen:

Glaubt man als Lehrkraft, seine Schüler und Schülerinnen auf zentrale Vergleichsarbeiten wie VERA8 vorbereiten zu müssen, kann auch auf Maßnahmen zurückgegriffen werden, die weniger auf den langfristigen Lernerfolg zielen als auf gute Testergebnisse. Gerade Lehrkräfte mit negativen Kontroll- und Kompetenzüberzeugungen (Typ B‘ und Typ S‘) müssen ggf. auf solche Maßnahmen zurückgreifen (Hypothese 2.3). Dazu zählen sicherlich die entsprechenden Test Wiseness-Strategien, wenngleich die Aufgaben von VERA8 kaum für Tricks & Tipps dieser Art anfällig scheinen, soweit diese Strategien nicht auch die Testvalidität erhöhen.

Die Befunde zeigen aber genau den gegenteiligen Effekt. Lehrkräfte der Typen A’ und G’ scheinen eher dazu geneigt zu haben, ihren Schülerinnen und Schülern auch Strategien zu

vermitteln, die eher zu richtigen Lösungen führen, auch wenn die Schüler und Schülerinnen die zum Lösen benötigte Kompetenz gar nicht besitzen.

Zur Hypothese H2.4 – geringes Engagement der Lehrkräfte führt zu Schülerzentrierung:

Die größte Gruppe bilden in dieser Studie Lehrkräfte, die sich durch positive berufliche Emotionen auszeichnen und dabei gleichzeitig mittelmäßige bis schlechte Werte in den Bereichen Arbeitsengagement sowie Kontroll- und Kompetenzüberzeugungen aufweisen. Dieses als Typ S‘ (=Schontyp) bezeichnete Muster stellt eine Konstellation dar, die ein gutes Haushalten mit geringen Ressourcen annehmen lässt. In Bezug auf eine Vorbereitungsphase scheinen dazu zwei Optionen zu passen: die Vorbereitung so weit wie möglich in die Verantwortung der Schüler und Schülerinnen zu verlagern und die Vorbereitung mit bereits konzipiertem Material wie Vorbereitungsheften zu gestalten (Hypothese 2.4).

Es zeigte sich in der Tat, dass Lehrkräfte des Typs S‘ selten in ihrem Unterricht eine ausführliche Übungs- und Wiederholungsphase mit allen Inhaltskompetenzen durchgeführt haben. Deutlicher ist der Unterschied sogar noch bei den Prozesskompetenzen, obwohl diese generell einen niedrigeren Stellenwert haben. Erwartungskonform gaben Lehrkräfte des Typs S‘ häufiger an, die Wiederholung ihren Schülerinnen und Schülern (nur) empfohlen zu haben. Dazu passt weiterhin, dass aus dieser Gruppe zwei Drittel der Lehrkräfte eine außerschulische Vorbereitung nicht nur empfohlen haben, sondern sie wahrgenommen haben, dass sich in zwei Drittel der Klassen die Schülerinnen und Schüler tatsächlich auch entsprechend vorbereitet haben, während jenes in den drei anderen Gruppen ein geringerer Anteil wahrgenommen hat. Der Annahme folgend nutzten die Lehrkräfte des Typs S‘ auch häufiger als andere Kompetenz- und vor allem Vorbereitungshefte, und zwar ohne damit jeweils eine Wiederholung aller Inhalts- und Prozesskompetenzen im Blick zu haben. Hier sind die Unterschiede aber statistisch unbedeutend. Jene Lehrkräfte ließen zum Teil aber ausgewählte Prozesskompetenzen üben. Darüber hinaus zeigte sich gerade für diese Lehrkräfte, dass sie zwar bei der Nutzung von Vorbereitungs- oder Kompetenzheften nicht mehr Unterrichtsstunden für die Vorbereitung aufwendeten, es ihnen aber gelang, die Zeit variantenreicher zu gestalten.

Speziell für Lehrkräfte des Typs S‘ erweisen sich die von den Schulbuchverlagen angebotenen Materialien als sinnvolle Unterstützung. Die für die Vorbereitung reservierte Unterrichtszeit wird mit Hilfe dieser Hefte, nicht umfangreicher, aber effektiver gestaltet und bietet für die Schüler und Schülerinnen dieser Lehrkräfte eine größere Wahrscheinlichkeit, eine qualitativ bessere Vorbereitung zu erleben. Die Steuerungswirkung zentraler Vergleichsarbeiten entfaltet sich für diese Lehrkräfte daher eher mittelbar.

Zu den Hypothesen H2.5 & H2.6 – keine Unterschiede in Vorbereitungsveränderung:

Der Umfang und die Variation in einer Vorbereitungsphase sollten u.a. an den Stellenwert der Leistungsmessung geknüpft sein. In Anbetracht der eher wenig positiven Befunde bzgl. der Nutzung und Reflexion von Ergebnissen aus zentralen Vergleichsarbeiten (Groß Ophoff, 2013; Hosenfeld, 2010; Koch, 2011) kann man zu dem Schluss kommen, dass der Vorbereitungsumfang und die Variation abnehmen sollten. Besonders diejenigen, die nur geringe Ressourcen zur Verfügung haben, sollten auf eine geringe Nutzung mit einem reduzierten Aufwand reagieren (Hypothesen H2.5 & H2.6).

Die Selbsteinschätzung der befragten Lehrkräfte konnte diese Annahme nicht bestätigen. Zwar haben in den Gruppen Typ S‘ und Typ B‘ weniger Lehrkräfte wahrgenommen, im Jahr 2010 mehr vorbereitet zu haben als bei den Malen davor, jedoch sind die Unterschiede minimal. Zudem hat sich nach Wahrnehmung der meisten Lehrkräfte die Intensität der Vorbereitung nicht verändert. Letzteres gilt für alle vier Gruppen in gleichem Maße.

Eine Erklärung für diesen Befund könnte in der tatsächlich beigemessenen Bedeutung von VERA8 liegen. Lehrkräfte des Typs S‘ maßen nach eigener Wahrnehmung VERA8 durchschnittlich minimal mehr Bedeutung zu als sie das von ihren ebenfalls betroffenen Kollegen und Kolleginnen annahmen. Es ist daher nur konsequent, dass sich (möglicherweise) auch die Vorbereitungsintensität erhöhte. Für Lehrkräfte des Typs B‘ kann aber zumindest die Hypothese H2.6 eingeschränkt bestätigt werden. Lehrkräfte aus dieser Gruppe wiesen VERA8 einen geringeren Stellenwert zu als andere Lehrkräfte. In der Gruppe hingegen wurde vor allem das Interesse der eigenen Schulleitung an VERA8 als besonders hoch eingestuft. Auch dies scheint ein nachvollziehbarer Grund für eine erhöhte Intensität der Vorbereitung zu sein, wenn man unterstellt, dass die Schulleitung diese Kolleginnen und Kollegen vermehrt unter Druck setzte. In Bezug auf den Stellenwert können die getroffenen Annahmen daher nicht bestätigt werden. Die mehrfach getätigten Befunde einer geringen Nutzung und Reflexion (zweite innovierende Funktion) sind folglich losgelöst von einer zunehmenden Intensität der Vorbereitung auf VERA8 (eingeschränkt erste innovierende Funktion).

8.1.3 Testcoaching als Reaktion auf (angekündigtes) Feedback

Zu F3: Inwiefern lassen sich Unterschiede zwischen Lehrkräften in Umfang und Qualität der Vorbereitung durch die Typisierung der Lehrkräfte nach Einstellungen zu und Umgang mit Rückmeldungen aus den Vergleichsarbeiten erklären?

Der zweite differentielle Ansatz verbindet die durchgeführte Vorbereitungsphase mit einer Einteilung in Feedback-Reflexions-Typen. Grundlage dieser Überlegungen war die Feedback-Interventions-Theorie von Kluger und DeNisi (1996). Nach dieser ist die Auseinandersetzung

mit angebotenem Feedback umso größer, je mehr Ressourcen in den Bereichen Selbstwirksamkeitserwartung, Gewissenhaftigkeit und erlebte Unterstützung vorhanden sind, die Ergebnisse mit dem eigenen Selbst attribuiert werden und eine Internalisierung der Ziele stattgefunden hat. Zusätzlich wurden Reflexionstypen angenommen, die aus Befunden von Stamm (2003) und Hosenfeld (2010) hervorgingen. Letztere wiederum basieren auf dem Rahmenmodell der pädagogischen Nutzung von Vergleichsarbeiten nach Helmke & Hosenfeld (2005) mit den Stufen Rezeption, Reflexion, Veränderung (und Evaluation).

VERA8 wurde hierbei als angebotenes Feedback für Lehrkräfte aufgefasst. Die Grundannahmen lauteten, (a) dass sich in der Nutzung vergangener VERA8-Ergebnisse und der beabsichtigten Nutzung von VERA8-Ergebnissen zeigt, inwieweit VERA8 als angebotenes Feedback wahrgenommen wurde, und (b) dass die jeweils praktizierte Vorbereitungsphase eine Reaktion der Lehrkräfte auf vorheriges Feedback oder das angekündigte Feedback war. Je nach Erfahrung mit vergangenen VERA8-Durchgängen und je nach Erwartung über die anstehende Rückmeldung sollten Lehrkräfte die Genese der Rückmeldung unterstützt oder zu ihren Gunsten beeinflusst haben.

Auch für diese zweite Grundannahme wurden für verschiedene Klassenmodelle latent Klassenanalysen durchgeführt. Als das beste Klassenmodell erschien das Modell M22. Dieses berücksichtigt einerseits als mögliche verfügbare Ressourcen die Selbstwirksamkeitserwartung [SWE], die Gewissenhaftigkeit [GW] und die erlebte Unterstützung durch die Schulleitung [USL] sowie eine allgemeine Akzeptanz der Kernlehrpläne [KL], andererseits umfasst es Skalen zur Rezeption von vorherigen VERA8-Ergebnissen [REPZ], deren Reflexion [RFL] und die Unterrichtsveränderungsbereitschaft [VEA]. In diesem Klassenmodell wurde folglich die FIT und die Idee des Rahmenmodells von Helmke & Hosenfeld miteinander verbunden. Datengrundlage für diese Analysen war der Datensatz aus Studie B.

Auf Grundlage des Datensatzes aus Studie A wurde das Klassenmodell M21 berechnet. Dieses umfasst neben den Skalen zur Gewissenhaftigkeit und Selbstwirksamkeitserwartung die Attribution der Schülerleistung [LA] und drei Skalen zu den Prozesskompetenzen als Unterrichtsziele [ZieleArK, ZielePLM, ZieleW]. Die Ergebnisse der Latent-Class-Analyse wiesen für die Vier- und Fünf-Klassen-Lösung eine ausreichende Modellpassung auf, ließen sich aber inhaltlich schwierig interpretieren. Die Datengrundlage bot dadurch entgegen der vorausgegangenen Annahme keine Möglichkeit, Lehrkräfte ausschließlich aufgrund der FIT zu klassifizieren.

Daher wurde für die zweite differentielle Analyse mit dem Klassenmodell M22 die Vier-Klassen-Lösung als die beste berücksichtigt. Die vier Muster (Typ a, Typ b, Typ nI & Typ nII) lassen sich dabei sowohl mit den theoretisch in Kap. 5 hergeleiteten Feedback-Typen assoziieren als auch mit den drei Reflexionstypen von Hosenfeld verbinden. Das Muster Typ a (in Anlehnung an die Klassifikation von Stamm als „Alibi“-Nutzer bezeichnet) bildeten knapp dreißig Prozent der klassifizierten Lehrkräfte, die sich durch mittlere Werte im Bereich der Ressourcen und durch eine geringe Akzeptanz der Kernlehrpläne auszeichneten (daher

auch Feedbacktyp F21.D) und sich nur stark eingeschränkt mit den Ergebnissen aus VERA8 auseinandergesetzt hatten, obwohl sie gleichzeitig eine größere Veränderungsbereitschaft angaben. Ein weiteres Viertel der Lehrkräfte wurde als Muster Typ b (entsprechend von „Blockierer“) gruppiert. Hierin sind Lehrkräfte des Feedbacktyps F21.C zusammengefasst, die entsprechend über wenige Ressourcen im Bereich Selbstwirksamkeitserwartung und Gewissenhaftigkeit verfügten und die curricularen Vorgaben nur eingeschränkt akzeptierten. Obwohl sie gleichzeitig teilweise durchaus eine Unterstützung durch die Schulleitung wahrnahmen, partizipierten sie am wenigsten am Schema von „Rezeption-Reflexion-Veränderung“. Beide Klassen bildeten zusammen den Rezeptionstyp R22.C. Demgegenüber zeichneten sich die beiden anderen Klassen durch Lehrkräfte aus, die sehr wohl intensiver mit den Ergebnissen aus VERA8 umgegangen sind. Die eine Gruppe (Typ nI – in der Studie B nur gut zehn Prozent der Lehrkräfte) verfügte dabei über durchweg gute bis sehr gute Ressourcen und akzeptierte die curricularen Vorgaben zum Unterricht voll und ganz (Feedbacktyp F21.A). Lehrkräfte dieses Typs kamen aber nur bis zur Stufe der Rezeption oder Reflexion, da es für sie annehmbar keinen Veränderungsbedarf gab (Reflexionstyp R22.B). Lehrkräfte der Klasse Typ nII verfügten nur eingeschränkt über notwendige Ressourcen und zeigten nur eine durchschnittliche Akzeptanz der Kernlehrpläne (größte Übereinstimmung mit Feedbacktyp F21.B). Über ein Drittel der Lehrkräfte gehörten zu dieser Klasse. Sie besaßen aufgrund der geringeren, aber vorhandenen Ressourcen durchaus ein größeres Potenzial, auch Veränderungen ihres Unterrichts anzustreben (Reflexionstyp R22.A).

Im ausgewählten Klassenmodell sind die Konstrukte der Klassifikation in Feedbacktypen und der Klassifikation in Reflexionstypen kombiniert worden. Dass sich dieses Modell als das beste Modell im zweiten differentiellen Ansatz herausstellte, belegt die vorher angenommene Verbindung der Bedingungsebene (Feedback-Interventions-Theorie) und der Handlungsebene (Rahmenmodell von Helmke & Hosenfeld). Nur die Kombination beider Ansätze ließ eine sinnvolle Klassifizierung zu (Forschungsfragen F3.1 und F3.2). Es führt aber andersherum dazu, dass die in Kap. 5 formulierten Annahmen kombiniert betrachtet werden müssen. Außerdem fand die Leistungsattribution in dem gewählten Modell keine Berücksichtigung.

Zu den Hypothesen H3.1 bis H3.5 – FIT sagt Vorbereitungsverhalten in Teilen richtig voraus:

Die Grundannahme für den zweiten differentiellen Ansatz lässt sich in der Art spezifizieren, dass eine umfangreiche Vorbereitung nur dann praktiziert wird, wenn die Lehrkraft die Zielvorgaben einerseits internalisiert hat, sie aber befürchtet, die Zielvorgaben nicht zu erreichen. Die Vorbereitung dient dann dazu, eine antizipierte Lücke zu vermeiden (als Pro-Aktion) oder ein in vorherigen Durchgängen identifiziertes Defizit zu beheben (als Re-Aktion). Andersherum wurde angenommen, dass eine umfangreiche Vorbereitung nicht rational erscheint, wenn entweder die Ziele nicht internalisiert sind oder auch ohne

Vorbereitung gute Ergebnisse zu erwarten sind. VERA8 musste zusätzlich gleichzeitig auch als angebotenes Feedback wahrgenommen werden.

Die Befunde aus der Studie B können diese Annahmen teilweise für den Umfang der Vorbereitung bestätigen. Lehrkräfte des Typs nII haben ihre Schülerinnen und Schüler durchschnittlich mehr als zwei Stunden länger vorbereitet als Lehrkräfte des Typs b und des Typs nI. Dies ist ein deutlicher Unterschied und der Unterschied ist größer als die Unterschiede zwischen den Typen in der Studie A unter dem Klassenmodell M11a. Gleichzeitig beträgt der Vorbereitungsumfang der Lehrkräfte aus dem Muster Typ a aber ebenfalls über acht Unterrichtsstunden und weist damit ebenfalls zumindest einen Unterschied von mehr als einer Unterrichtsstunde zu den Mustern Typ b und Typ nI auf.

Der zweite Indikator für die Intensität der Vorbereitung, die Anzahl der FA-Maßnahmen, unterstützt den Befund zum Vorbereitungsumfang. Lehrkräfte des Typs nII nutzen als einzige mehrheitlich drei Maßnahmen, während die anderen mehrheitlich nur zwei Maßnahmen umsetzten, und mit jeder vierten Lehrkraft aus dieser Gruppe nutzten auch deutlich mehr Lehrkräfte vier oder mehr Maßnahmen als aus jeder anderen Gruppe. Demgegenüber zeigt das Muster Typ nI wiederum das umgekehrte Bild und Lehrkräfte nutzten hier die wenigsten FA-Maßnahmen. Es war folglich zutreffend, dass Lehrkräfte, die sowohl die Schülerleistungen in ihrer Verantwortung verorten als auch die Unterrichtsvorgaben akzeptieren, aber eine niedrige SWE besitzen (Typ F21.B), am umfangreichsten vorbereiteten (Hypothesen H3.3 – und auch H3.4).

Festgehalten werden muss an dieser Stelle aber, dass trotz der deutlichen Unterschiede zwischen den Mustern für alle vier Muster die Anzahl der durchschnittlich aufgewendeten Unterrichtsstunden erneut die Zahl von bis zu zwei Unterrichtsstunden übersteigt. Die Umkehrung (Hypothese H3.2), Lehrkräfte, die sowohl die Schülerleistungen in ihrer Verantwortung verorten als auch die Unterrichtsvorgaben akzeptieren und eine hohe SWE besitzen (Typ F21.A), bereiteten minimal vor, erwies sich als nicht zutreffend. Auch in Studie B zeigte sich daher eine Steuerungswirkung von VERA8, die in dieser Weise nicht intendiert ist.

Der mittelbare Steuerungseffekt über Vorbereitungs- und Kompetenzhefte ist vor allem für Lehrkräfte des Typs b geringer als für die anderen Lehrkräfte. Von den „Blockierer“ nutzt nicht einmal jeder zweite eine der beiden Heftarten in der Vorbereitungsphase, in den anderen Gruppen sind es jeweils zwei Drittel. Auch in der Gruppe Typ nI greifen Lehrkräfte folglich mehrheitlich auf diese Hefte zurück. D.h. einerseits gelangt (ebenso) auf diesem Weg eine neue Aufgabenkultur in den Unterricht auch bei Lehrkräften, die wahrscheinlich bereits einen sehr guten Unterricht anbieten, andererseits sieht sich offensichtlich auch ein großer Teil dieser Lehrkräfte genötigt, auf solche Hefte zurückzugreifen. Die Hypothese H3.5, Lehrkräfte, die dem Typ R.22.B und Typ R.22.C zugeordnet wurden, sollten minimal vorbereitet haben, muss daher erneut zurückgewiesen werden. Auch Hypothese H3.1 ist nicht vollständig haltbar und muss als zu unterkomplex zurückgewiesen werden.

Zu Unterschieden in der Qualität der Vorbereitungshase

Der Blick auf die einzelnen Maßnahmen bringt zusätzlich Erkenntnisse über Unterschiede in der Qualität der Vorbereitungsphase. Lehrkräfte der Gruppe Typ nII hatten eine größere Tendenz als andere Lehrkräfte, mit alten Aufgaben direkt im Unterricht vorzubereiten. Sie haben auch häufiger als andere VERA8-ähnliche Aufgaben in vorherige Klassenarbeiten eingebaut. Letztere Maßnahme wurde auch überdurchschnittlich häufig von Lehrkräften des Typs a angewendet. Statt jeder zweiten Lehrkraft war es hier aber nur jede dritte. Lehrkräfte des Musters Typ nI zeichneten sich bzgl. der FA-Maßnahmen hingegen vor allem dadurch aus, dass sie die Vorbereitung in die Hände der Schülerinnen und Schüler legten, indem sie beispielsweise auf die offizielle Internetseiten verwiesen. Die Integration von VERA8-ähnlichen Aufgaben in vorherige Klassenarbeiten ist für Studie A derart interpretiert worden, dass VERA8 eine neue Aufgabenkultur auch in den vorhergehenden Unterricht implementiert. Die erste innovierende Funktion kam demzufolge nur bei Lehrkräften der Typen nII & a sichtbar zum Tragen.

Je mehr man VERA8 als angebotenes Feedback versteht, desto mehr sollte ein Familiarity Approach im Fokus stehen, welches eine höhere Testvalidität als Ziel hat. Andersherum ist es zumindest für die Themen (q), wie man den Aufgabenstellungen der VERA8-Aufgaben die wichtigen Informationen entnimmt, und ggf. auch (p), wie man die Aufgabenstellung der VERA8-Aufgaben richtig versteht, schwieriger als bei anderen FA-Maßnahmen, Lernerfolge sichtbar werden zu lassen. Lehrkräfte, die die Vorbereitung vor allem auf ein gutes Abschneiden ausrichten, sollten eher zu anderen Maßnahmen greifen, die sich konkreter mit den zu erwartenden Inhalten beschäftigen. In Bezug auf die Frage, inwieweit Lehrkräfte den besonderen Aufgabentypus von VERA8 angesprochen haben, zeigte sich vor allem für das Muster Typ b ein erwartungskonformes Bild. Lehrkräfte dieses Typs haben wesentlich seltener die drei Themen zum Aufgabentypus angesprochen und sich damit weniger um eine höhere Testvalidität bemüht. Dies entspricht der Annahme, dass diese Lehrkräfte VERA8 am wenigsten als Feedback für ihren Unterricht annehmen wollten.

Für die Übungs- und Wiederholungsphasen zeigte sich ebenfalls eine Sonderstellung der Klasse Typ b. Die Vorbereitungsphase wurde von diesen Lehrkräften seltener dafür genutzt, alle Inhaltsbereiche zu wiederholen oder üben zu lassen, vor allem aber gilt dies mehr noch für die Prozesskompetenzen. Während in den drei anderen Gruppen mindestens jede zweite Lehrkraft auch die Prozesskompetenzen integrierte, spielten diese für Lehrkräfte des Typs b nur in jedem vierten Fall überhaupt bewusst eine Rolle. Inhaltlich verfehlt VERA8 für Lehrkräfte dieser Art einen der wichtigsten Aufträge.

Insgesamt ergaben sich für die Klasse nII nur hohe Werte beim zeitlichen Umfang der Vorbereitung und der Anzahl der durchgeführten FA-Maßnahmen, nicht aber im Bereich der inhaltlichen Gestaltung. Lehrkräfte des Typs b hingegen zeichneten sich dadurch aus, dass sie insgesamt die geringste Vorbereitungsintensität zeigten. Sie haben die wenigste Unterrichtszeit in die Vorbereitung investiert, die wenigsten FA-Maßnahmen durchgeführt und (schon zwangsläufig) auch inhaltlich die wenigsten Themen in der Vorbereitung

behandelt. Lehrkräfte des Typs nI haben ebenfalls in geringerem zeitlichen Rahmen vorbereitet. Sie gaben aber mehrheitlich an, trotzdem die drei Bereiche der Aufgabestellung thematisiert und eine Vorbereitung durchgeführt zu haben, die auch eine Wiederholung der Inhaltskompetenzen und (wie insgesamt für alle Lehrkräfte geltend) auch der Prozesskompetenzen als Ziel hatte. Die Klasse Typ a wiederum war durch deutliche Parallelen zur Klasse Typ nII gekennzeichnet: Diese sind vor allem ein ähnlich hoher zeitlicher Vorbereitungsumfang, aber auch nur minimale Differenzen zu den Werten in den anderen Analyseabschnitten.

8.2 Reflexion des Untersuchungs- und Analysevorgehens

Bevor abschließend die Ergebnisse dieser Arbeit noch einmal in einer Gesamtdiskussion mit Blick auf die Perspektive für Praxis und Forschung gewürdigt werden, sollen vorab wesentliche Einschränkungen der Gesamtergebnisse dieser Arbeit diskutiert werden:

(1) Inwieweit ist das in den beiden Studien erhobene Bild über das Vorbereitungsverhalten insgesamt repräsentativ?

Wie in Kapitel 6 beschrieben, handelt es sich in beiden Studien und in allen vier Teilstudien in der Realisierung um Gelegenheitsstichproben (Weick, 1976). Obwohl zumindest das in allen vier Teilstudien erhobene Vorbereitungsverhalten aggregiert als Vollerhebung angelegt war, konnte erwartungsgemäß keine entsprechende Rücklaufquote realisiert werden. Die Rücklaufquoten liegen allerdings im Rahmen vergleichbarer Studien. Auch der Vergleich der Altersangaben aus den realisierten Stichproben und den amtlichen Statistiken zeigt zwar Abweichungen, diese lassen sich aber durch die Personalpolitik der Schulen sinnvoll erklären. Der Vergleich zwischen den vier Teilstudien zeigt insgesamt ein homogenes Bild des Vorbereitungsverhaltens. Es ist unwahrscheinlich, dass zufällige Verzerrungen viermal zu einem ähnlichen Bild führen. Insgesamt muss daher nicht zwingend von einer zufälligen Verzerrung ausgegangen werden.

Möglich ist aber trotzdem eine systematische Verzerrung in der Weise, dass sich eventuell unabhängig von den zur Repräsentativitätsprüfung herangezogenen Merkmalen gerade diejenigen Lehrkräfte zu einer Teilnahme an der jeweiligen Studie entschlossen haben, die ein bestimmtes (ggf. umfangreiches) Vorbereitungsverhalten zeigten. Beachtet werden muss in diesem Zusammenhang, dass eine größere Ausschöpfung in beiden Studien jeweils zu einem leicht geringeren Stundenvolumen bei der Vorbereitung führte. Zwar sind die Unterschiede nur in einem Fall signifikant, die Teilstudien mit gelben Haftzetteln bilden die tatsächliche Vorbereitung möglicherweise aber besser ab. Es kann daher nicht

ausgeschlossen werden, dass sich gerade diejenigen Lehrkräfte an den Studien beteiligten, die intensiver vorbereiteten.

(2) Inwieweit sind die Effekte bedeutsam?

Alle für die verschiedenen Gruppenvergleiche berichteten normierten Effektstärken sind ausschließlich kleine bis mittlere Effekte. Als groß gelten Effekte in den Sozialwissenschaften erst ab einem Wert von r=.371 (Volker, 2006). Dies kann in der Tat derart gedeutet werden, dass die dargestellten Unterschiede von geringer Bedeutung sind. Hierfür spricht die grundsätzliche Ausprägung der einzelnen Vergleichsvariablen wie beispielsweise die Anzahl der aufgewendeten Unterrichtsstunden für die Vorbereitung. In allen Gruppen der differentiellen Analysen sind die Ausprägungen ähnlich hoch. Zumindest das angenommene theoretische Modell im zweiten differentiellen Ansatz beinhaltet aber nur Faktoren, die auch bei der Gruppeneinteilung berücksichtigt wurden. Dem Modell folgend müssten sich die Merkmalsausprägungen deutlicher voneinander unterscheiden.[137]

Das genutzte Effektmaß darf aber nicht nur im Kontext der Fachdisziplin interpretiert werden, sondern muss auch den speziellen Kontext der Studien berücksichtigen. Die dieser Arbeit zugrunde liegenden Studien untersuchen keine singulären Unterschiede zwischen zwei eindeutig identifizierbaren Gruppen. Die Gruppeneinteilung an sich ist eine Projektion eines – hier vierdimensionalen – Klassifikationsvektors auf seine jeweils stärkste Ausprägung. Dabei sind die Ausprägungen Wahrscheinlichkeitswerte. Jede Lehrkraft gehört mit einer geringeren Wahrscheinlichkeit auch in die jeweils anderen Gruppen. Die berechneten Effekte sind daher als minimale Untergrenze zu deuten. Trotzdem sind die Unterschiede zwischen den Gruppen in der Tat tendenziell gering.

(3) Inwieweit sind die Ergebnisse auf andere Bundesländer, Fächer und Schulformen übertragbar?

Beide in dieser Arbeit berücksichtigten Studien haben nur das Vorbereitungsverhalten von Mathematik-Lehrkräften auf VERA8 in Nordrhein-Westfalen untersucht. Eine Übertragung der aggregierten Befunde auf andere Bundesländer und die anderen bei VERA8 berücksichtigten Fächer ist aus den Studien allein nicht seriös möglich. Mittlerweile ist die Teilnahme an VERA8 in einigen Bundesländern nur noch alternierend in einem der Fächer vorgeschrieben. Es muss angenommen werden, dass sich in diesen Bundesländern der Stellenwert von VERA8 auch bei den Lehrkräften durch diese Entscheidung verändert hat. Möglicherweise ist bzw. war der Stellenwert in Bundesländern geringer, die eine kürzere Tradition mit zentralen Vergleichsarbeiten haben und nicht den gleichen administrativen Unterstützungsgrad bieten wie er in Nordrhein-Westfalen angeboten wird. Die beiden

[137] Für den ersten differentiellen Ansatz gilt dies nicht in gleicher Weise, da dort nur ein Ausschnitt des angenommenen erweiterten Modells der Lehrer-Handlungskompetenz abgebildet werden konnte.

Studien zeigen, dass ein hoher Stellenwert von VERA8 und der Vorbereitungsumfang zusammenhängen. Bzgl. der anderen Fächer verhält es sich hingegen anders. Anders als in Mathematik werden für die sprachlichen Fächer Monate vor der Erhebung Schwerpunkte angekündigt. Durch die Schwerpunktsetzung tendieren Lehrkräfte dazu, in diesen intensiver vorzubereiten als in den anderen Bereichen, wie die Interviewstudie zeigte (Hahn, 2008). Es ist daher für die sprachlichen Fächer eher zu erwarten, dass die Vorbereitungsintensität zunimmt. Vor allem aber sollten die Übungs- und Wiederholungsphasen durch Phasen gezielten Übens mit Blick auf die Schwerpunkte ersetzt werden. Auch lassen die Ergebnisse nicht zwingend ein ähnliches Vorbereitungsverhalten für Mathematik in anderen Schulformen (z.B. Grundschulen) erwarten. Mathematik wird in Schulformen häufiger fachfremd unterrichtet. Das fachfremde Unterrichten und die andere Schulform haben voraussichtlich Auswirkungen auf das Selbstbild der Lehrkräfte, deren Auswirkungen im Rahmen der Selbstbild-Theorien nicht vorhersehbar sind.

Die beschriebenen Einschränkungen treffen aber nicht im gleich starken Maße auf die differentiellen Ansätze zu. Die beiden analysierten Gruppeneinteilungen wurden in beiden Studien jeweils auf Grund allgemeiner arbeitspsychologischer Modelle (für Lehrkräfte) vorgenommen und sind daher keine fachspezifischen oder vom Bundesland oder der Schulform abhängigen. Von diesem Standpunkt aus können alle Ergebnisse der differenziellen Analysen auch auf andere an VERA8 beteiligte Lehrkräfte übertragen werden.

8.3 Fazit und Perspektive für Praxis und Forschung

8.3.1 Fazit: Welche Schlüsse lassen sich aus der beobachteten Sachlage über die Steuerungswirkung zentraler Vergleichsarbeiten ziehen?

Zu Beginn dieser Arbeit wurden steuerungs- und lerntheoretische Ziele (nach Kühn, 2010, Maritzen, 2008, und Tresch, 2008, bzw. nach Heymann, 2007, und Parveva et al., 2009) benannt. Ein Abgleich dieser Ziele und den vorliegenden Befunden führt zu den angestrebten Erkenntnisse über die Steuerungswirkung zentraler Vergleichsarbeiten. Mit der ebenfalls in Kapitel 2 skizzierten Educational Governance-Perspektive und der dort beschriebenen Perspektive der Schulqualitätsforschung gelangt diese Arbeit abschließend zu Erklärungen für die dargelegten Befunde:

Die vorliegende Arbeit hat die tatsächliche Wirksamkeit der berichteten Vorbereitungsphasen nicht untersucht, sodass eine Verzerrung der Testergebnisse nicht zwingend erfolgt sein muss. Es ist nicht ausgeschlossen, dass all die investierte Mühe der Lehrkräfte sowie der Schülerinnen und Schüler wirkungslos in Bezug auf das tatsächliche Testergebnis war. Vor dem Hintergrund, dass in den Vorbereitungsphasen nicht wesentlich

anderes gemacht wurde als im Alltag vor Klassenarbeiten, erscheint dies allerdings unwahrscheinlich. Auch die in den Kapiteln 2 & 3 berichteten Befunde zu anderen Formen der standardisierten Testverfahren lassen eher eine Wirksamkeit der Vorbereitungsphasen annehmen, zumindest in der Form, dass eine Vorbereitung auf VERA8 einen kurzfristigen Lernzuwachs ermöglicht. Es steht somit in Frage, ob mit zentralen Vergleichsarbeiten in der vorliegenden Form eine Qualitätssicherung oder gar -erweiterung betrieben werden kann. Die Idee der Qualitätssicherung und -erweiterung beruht auf der Annahme, dass eine Output-Kontrolle in Form von zentralen Vergleichsarbeiten zu einer höheren Gesamtprozess-Qualität führe. Im Zusammenhang mit zentralen Vergleichsarbeiten wie VERA8 konnten die Studien dieser Arbeit keine Belege dazu liefern. Stattdessen muss aus den beiden Studien gefolgert werden, dass Lehrkräfte mehrheitlich selbst diesen Zusammenhang nicht wahrnehmen und sich daher in der Pflicht sehen, eine spezielle Vorbereitungsphase einzubauen. Die Realisierung der deskriptiven Funktion ist daher ebenso zweifelhaft. Gleichzeitig werden im Fall der Wirksamkeit spezieller Vorbereitung auch die Ziele (b) wissenschaftliche Erkenntnis über Wirksamkeit schulischer Arbeit zu erlangen, (e) Unterstützung einzelner Schulen zu ermöglichen, (g) über die Leistungen der Einzelschulen zu informieren und (h) ggf. Wettbewerb zu ermöglichen unterlaufen.

Unabhängig von der tatsächlichen Wirksamkeit der Vorbereitung reicht aber bereits die Annahme der Lehrkräfte über die Wirksamkeit einer speziellen Vorbereitungsphase aus, um zwei weitere Ziele auszuhebeln. Wenn die Ergebnisse aus zentralen Vergleichsarbeiten genutzt werden sollen, um (c) Impulse für die Schulentwicklung anzuregen und (d) Lehrkräften und Schulen ein Feedback anzubieten, müssen die Schulen und Lehrkräfte wahrnehmen, dass die Ergebnisse ein korrektes Bild des Unterrichtserfolgs darstellen. Bei der Konzeption von zentralen Vergleichsarbeiten muss somit die Beweislast erbracht werden, dass die Testergebnisse durch ein geändertes Unterrichtsangebot beeinflusst werden können. Die vorliegenden Befunde über die bei fast allen befragten Lehrkräften vorgefundenen Vorbereitungsphasen sprechen aber zusätzlich dafür, dass VERA8 als Vergleichsarbeit mit der Sozialnorm als Referenzrahmen in der Wahrnehmung der Lehrkräfte den tatsächlichen Unterrichtserfolg nicht abbildet.

Hierzu lieferten die beiden differentiellen Ansätze wichtige Erkenntnisse: Lehrkräfte mit geringen notwendigen Ressourcen (beispielsweise als Typ B‘ klassifizierte), die von ihren Schülerinnen und Schülern geringere Leistungen erwarten können, verzerren die Testergebnisse möglicherweise bewusst durch eine spezielle Vorbereitung. Lehrkräfte mit großen notwendigen Ressourcen aber (beispielsweise als Typ G‘ klassifizierte), die eigentlich überdurchschnittliche Ergebnisse von ihren Klassen erwarten können, bereiten ihre Klassen ebenfalls und intensiver vor. Während sich einerseits unabhängig von qualitativen Unterschieden in den personenbezogenen Ressourcen der Lehrkräfte eine hohe Vorbereitungsintensität zeigte, sind andererseits abhängig von den personenbezogenen Ressourcen Unterschiede in der Qualität der Vorbereitungsphase erkennbar gewesen. Es tritt offenbar auch hier der im Bildungsbereich häufig anzutreffende Matthäus-Effekt auf.

Zusätzlich nehmen Lehrkräfte möglicherweise an, eine spezielle Vorbereitung werde vorausgesetzt. Im Grunde resultiert die Vorbereitung aus einem Pflichtgefühl heraus, dass Ähnlichkeit zur bürokratischen und zur professionellen Rechenschaftslegung aufweist. Hierfür spricht u.a. ebenfalls der immer noch hohe Durchschnittswert an für die Vorbereitung aufgewendeten Unterrichtsstunden bei denjenigen Lehrkräften, die als Nutzertyp nI klassifiziert wurden, sich gerade als an der pädagogischen Nutzung von Ergebnissen aus zentralen Vergleichsarbeiten interessiert gezeigt haben und bisher gleichzeitig keinen Anlass für eine Unterrichtsveränderung erfahren haben. Der zweite differentielle Ansatz konnte daher zeigen, dass es tendenzielle Unterschiede in der Vorbereitungsintensität in Abhängigkeit vom Grad, mit der eine Lehrkraft VERA8 als Feedback-Angebot nutzen möchte, gibt, dass der Grad aber nur einen minimalen Einfluss auf die Vorbereitungsintensität aufweist.

Zumindest für die in den Studien befragten Lehrkräfte stellt sich VERA8 daher insgesamt nicht in der notwendigen Weise als ein Instrument dar, mit dem Qualitätsentwicklung betrieben werden könnte. Die Vorbereitungsintensität ist mit dieser Absicht nicht vereinbar. Infolgedessen können zentrale Vergleichsarbeiten wie VERA8 auch nicht (f) von der Konzeption von Tests entlasten, denn Lehrkräfte können offenbar in VERA8 keinen ausreichenden Nutzen für die Qualitätsentwicklung erkennen, für die sie ggf. Tests konzipieren müssten. Lediglich durch die Übernahme einzelner Aufgaben in andere Klassenarbeiten mag eine Entlastung für einige Lehrkräfte stattfinden. Für diese Lehrkräfte erfüllt VERA8 ebenfalls zumindest die erste innovierende Funktion.

Weiter kann aufgrund der flächendeckenden Nutzung alter VERA8-Aufgaben oder als VERA8-ähnlich eingeschätzter Aufgaben angenommen werden, dass die Intention des Kernlehrplans und der Bildungsstandards eingeschränkt illustriert werden. Inwieweit dies konkret passiert, muss vor dem Befund der mehrheitlich nur auf die Wiederholung von Inhaltskompetenzen ausgerichteten Übungs- und Wiederholungsphasen weiter untersucht werden. Die Bedeutung der Prozesskompetenzen ist voraussichtlich zumindest bis zum Erhebungszeitraum der Studien nicht ausreichend vermittelt worden.

Die offenbar flächendeckend genutzten Vorbereitungs- und Kompetenzhefte haben hierbei sogar eine eher positive Rolle. Gleichzeitig scheint mit ihrer Nutzung aber auch das Steuerungsprinzip der output-orientierten Steuerung ad absurdem geführt. Wenn für die Lehrkräfte aus der Einführung zentraler Vergleichsarbeiten und dem Angebot der Schulbuchverlage an Vorbereitungs- und Kompetenzheften deren Nutzung obligatorisch erscheint, wurde die Input-Steuerung der administrativen Ebene durch eine Input-Steuerung der Schulbuchverlage ersetzt.

Das zugrunde gelegte Input-Prozess-Output-Modell zeigt sich auch hier als fragwürdig. Die Kontrolle des Outputs durch VERA8 führt eben nicht zu dem von administrativer Seite angestrebtem Prozess. Dies zeigt sich dann noch einmal unter der Educational Governance-Perspektive und der Perspektive der Schulqualitätsforschung:

Die **Educational Governance-Perspektive** mit der Prinzipal-Agent-Theorie leitet den Fokus auf das Element der Rechenschaftslegung und erklärt den Zusammenhang zwischen durchgeführter Vorbereitung auf VERA8 und der von Lehrkräften antizipierten Wirkung ihrer Vorbereitung. Aus der Perspektive der Prinzipal-Agent-Theorie kann output-orientierte Steuerung grundsätzlich nur über Instrumente wie zentrale Vergleichsarbeiten funktionieren, wenn das unterstellte Input-Prozess-Output-Modell als lineares Modell zutreffend ist. Es reicht dabei nicht aus, dass das Modell den Wirkmechanismus von Testleistungen korrekt abbildet, das Modell muss von den Agenten, den Lehrkräften, in gleicherweise als korrekt akzeptiert werden. Das von der Prinzipal-Agent-Theorie beschriebene Kontrolldefizit (Jensen & Meckling, 1976; Kieser & Ebers, 2006) bleibt aber durch eine Vorbereitungsstrategie erhalten, die in Kauf nimmt, auch Lernen zu befördern, welches sich nur *kurzfristig* in besseren Testergebnissen niederschlägt. Solange Lehrkräfte nachvollziehbar davon ausgehen können, Kompetenzdefizite ihrer Schüler und Schülerinnen durch eine spezielle Vorbereitung im Vorfeld von VERA8 kaschieren zu können, findet keine echte Rechenschaftslegung durch VERA8 statt. Das hierzu identifizierte Hauptproblem: Gerade die in anderen Studien gezeigte geringe Auseinandersetzung mit den Ergebnissen aus VERA8 ermöglicht genau diesen – möglicherweise sogar falsche – Eindruck beizubehalten. Erst eine intensivere Auseinandersetzung mit den angebotenen Ergebnissen aus VERA8 kann dazu führen, dass eine spezielle Vorbereitung nicht mehr als effektiv angesehen wird, weil eine durchgeführte differenziertere Ursachen- und Aufgabenanalyse die Testeigenschaften von VERA8 besser verstehen lässt. Bei der nur oberflächlichen Auseinandersetzung mit den Ergebnissen aus VERA8 durch die Lehrkräfte fällt eine ineffektive Vorbereitung wenig auf. Ebeneso verhindert die flächendeckende Umsetzung von Testcoaching im Vorfeld von VERA8 eine Auseinandersetzung mit den Effekten dieses Testcoachings.

Die **Perspektive der Schulqualitätsforschung** erklärt die inhaltliche Gestaltung der durchgeführten Vorbereitungsphase mit Schwerpunkten auf dem Üben mit alten VERA8-Aufgaben und auf der Wiederholung von Inhaltskompetenzen. Während der input-orientierten Steuerung boten Lehrpläne durch eine klare Ausrichtung auf (mathematische) Inhalte und Verfahren für Lehrkräfte eine eindeutige Orientierung. Das Zulassungsverfahren für Schulbücher in Nordrhein-Westfalen wirkte zusätzlich unterstützend. Durch den Wandel zur output-orientierten Steuerung wird für Lehrkräfte das von der Systemtheorie Luhmanns benannte Technologiedefizit sichtbar, Inhalt und Prozess als Teil des Unterrichts sind nun offen. Aus Sicht der Lehrkräfte sind Lücken entstanden, die es zu füllen gilt. Dazu greifen sie auf Bewährtes zurück, nämlich einmal auf die Transformation eines problem-orientierten Unterrichts in Übungs- und Wiederholungsphasen mit klarer Orientierung an Testsaufgaben und außerdem auf externe Angebote der Schulbuchverlage (hier: Vorbereitungs- und Kompetenzhefte). Die mit den Ergebnissen aus VERA8 angebotenen Informationen über die Kompetenzen der Schülerinnen und Schüler und dem damit verbundenen Unterrichtserfolg verringern als Produktwissen das ausgemachte Technologiedefizit des Unterrichtens nicht, sondern lassen es noch größer erscheinen. Zusätzlich muss auf Grundlage des

dokumentierten Vorbereitungsbildes davon ausgegangen werden, dass Lehrkräfte den Ergebnissen nur einen eingeschränkten Wahrheitsgehalt zusprechen und dem durch VERA8 gewonnenen Produktwissen weniger Bedeutung beimessen als von administrativer Seite vorausgesetzt.

Die von den meisten Lehrkräften durchgeführte spezielle Vorbereitungsphase lässt sich begründet als Versuch deuten, im Sinne der neo-institutionellen Perspektive den bisherigen Unterrichtsstil als Kerntechnologie vor der institutionellen Umwelt zu schützen. Indem es (scheinbar) ausreicht, in den Wochen vor VERA8 den Unterricht mehr auf den dort geforderten Output auszurichten, wird man als Lehrkraft in die Lage versetzt, den Unterricht in der ganzen Zeit davor unbeeinflusst zu gestalten. Als Beleg dafür kann der inkonsequente Umgang der Nutzer-Typen in Studie B gelten, aber genauso das Vorbereitungsverhalten der als Typ G‘ in Studie A klassifizierten Lehrkräfte. Dem stehen nur vereinzelt Lehrkräfte und Fachgruppen gegenüber, die von Unterrichtsveränderungen berichten, die den vorherigen Unterricht insgesamt betreffen.

8.3.2 Welche Handlungsoptionen lassen sich aus den Ergebnissen für die administrative Ebene des Bildungswesens ableiten und welche Forschungslücken sind noch zu schließen?

Aus dem hier Dargelegten ergeben sich drei mögliche Handlungsoptionen für die administrative Ebene: (1) Bezüglich spezieller Vorbereitung auf VERA8 braucht es mehr Aufklärungsarbeit durch die administrative Ebene. Das Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen hat in der Zwischenzeit reagiert und die Hinweise zu einem sinnvollen Rahmen der Vorbereitung auf der Homepage erweitert. Die dort gegebenen Hinweise erscheinen nachvollziehbar und greifen das intuitive Verständnis der Lehrkräfte auf, auf VERA8 vorbereiten zu wollen. Es bleibt abzuwarten, inwieweit diese Maßnahme erfolgreich ist. Andere Bundesländer haben auch hier noch Nachholbedarf. Hilfreich wären sicherlich auch klarere Belege dafür, dass eine spezielle Vorbereitung nicht die erwünschten Effekte hat. Dazu wären aber Studien mit experimentellem Design nötig.

(2) Wie in allen Studien zur Nutzung von Ergebnissen aus zentralen Vergleichsarbeiten erweist sich die tatsächliche Nutzung auch im Auge der hier vorliegenden Befunde als ausbaufähig. Dabei wurde gleichzeitig die Verbindung zwischen der Überlegung, speziell auf zentrale Vergleichsarbeiten vorzubereiten, und die noch defizitäre Nutzung des angebotenen Feedbacks hingewiesen. Mit einem besserem Verständnis der Möglichkeit, die Ergebnisse pädagogisch zu nutzen, einem sicheren Verständnis der Datenlage selbst und mit Unterstützungsangeboten für die Unterrichtsentwicklung wurden drei Anknüpfungspunkte benannt. Es scheinen Unterstützungsmaßnahmen von administrativer Seite sinnvoll, die diese drei Problemfelder bei der Nutzung reduzieren. Die Weiterentwicklung des Unterrichts

lässt sich möglicherweise durch weitere Fortbildungsangebote erwirken. Für den sicheren Umgang mit den Daten bietet sich ein Multiplikatorensystem an.

(3) Die massive Vorbereitung auf VERA8 ist trotz der geringen Unterschiede zwischen Lehrkräften mit unterschiedlichen Ressourcen und unterschiedlicher Nutzungsabsicht auch ein Resultat der „Funktionsüberfrachtung“, vor der bereits Kühle und Peek warnten (2007). Wenn Lehrkräfte das durch VERA8 gewonnene Produktwissen für die eigene Weiterentwicklung nutzen sollen, sich aber gleichzeitig der Forderung nach Rechenschaftslegung über ihren Unterrichtserfolg ausgesetzt sehen, müssen manche Lehrkräfte sich entscheiden, welcher Forderung sie eher nachkommen wollen. Messen Lehrkräfte der Forderung nach Rechenschaftslegung größere Bedeutung bei, greifen sie mit der speziellen Vorbereitung ihrer Schülerinnen und Schüler auf VERA8 wissentlich auf Mittel zurück, die den tatsächlichen Unterrichtserfolg gerade nicht abbilden. Das gewonnene Produktwissen verliert für die Lehrkräfte an Wert. Da hier die deskriptive, die diagnostische, die zweite innovierende und die entwickelnde Funktion der Rechenschaftslegung gegenüberstehen, scheint eine Trennung der beiden Zielrichtungen angebracht, wie sie bei der Entwicklung von Aufgaben bereits praktiziert wird. Es wird demnach zwischen „Aufgaben zum Lernen“ und „Aufgaben zum Leisten“ unterschieden (Büchter & Leuders, 2007). Für die pädagogische Nutzung des mit zentralen Vergleichsarbeiten zu gewinnenden Produktwissens ist keine jährliche Vollerhebung notwendig. Als landesweite Referenzrahmen könnten auch die Ergebnisse früherer Erhebungen genutzt werden (unter der Annahme eines realistischen Abbilds der Kompetenzen). Wollen Lehrkräfte und Einzelschulen ihren Unterricht mit standardisierten Tests evaluieren, könnte die administrative Ebene dazu Unterstützung anbieten, die sich vom Aufwand nicht von der bisherigen Durchführungspraxis unterschiede. Dabei bliebe gleichzeitig die Möglichkeit gegeben, über durchgeführte Evaluationsverfahren von den Einzelschulen Rechenschaft zu verlangen. Parallel dazu könnte in größeren Abständen eine Leistungsmessung mittels einer Zentralstichprobe erfolgen, um die entwickelnde Funktion zu verwirklichen. Auch Groß Ophoff (2013) diskutiert diese Variante der Zentralstichprobe. Voraussetzung hierbei wäre aber, dass das zu erwartende Vorbereitungsverhalten der davon betroffenen Lehrkräfte Berücksichtigung fände und mögliche Effekte einer speziellen Vorbereitung einbezogen würden. Auch hierzu muss die Wirkung der speziellen Vorbereitung vorab untersucht werden. Mit den so über die Vorbereitungseffekte gewonnenen Ergebnissen ließen sich die Testergebnisse aus der Zentralstichprobe sinnvoll einordnen. Dazu bietet sich das in Kapitel 3 skizzierte lineare Modell an, aber auch eine multiplikative Berücksichtigung des Testcoachings wäre denkbar.

Abseits von der tatsächlichen Wirkung von Testcoaching im Rahmen von VERA8 ist vor allem die Untersuchung des Vorbereitungsverhaltens aus der in dieser Arbeit ausgesparten **Perspektive der Schulentwicklungsforschung** interessant. In diesem Zusammenhang könnte ein genauerer Blick auf die Frage lohnen, ob die Vorbereitungsintensität tatsächlich nach dem ersten Durchgang in der Regel zunimmt und ob es sich dabei um eine einmalige Zunahme handelt oder um eine kontinuierliche Steigerung. Der signifikante Unterschied

zwischen Lehrkräften ohne jegliche Erfahrung mit VERA8 und Lehrkräften mit VERA8-Erfahrung ist hierfür nur ein erstes Indiz. Weiter bietet auch ein möglicher Zusammenhang zwischen einer vorgenommenen allgemeinen Unterrichtsentwicklung und dem speziellen Vorbereitungsverhalten Forschungspotenzial. Die Befunde aus Studie A lassen hierzu erwarten, dass gerade Lehrkräfte mit gutem Unterricht auch qualitativ höher vorbereiten. In Studie A ist dieser Zusammenhang zwar unterstellt, er wird über die personenbezogenen Ressourcen aber nur mittelbar erfasst. Für die Messung der Unterrichtsqualität wären weitere methodische Zugänge nötig, beispielsweise Unterrichtsbeobachtungen. Diese könnten auch genutzt werden, um die Qualität der speziellen Vorbereitungsphasen noch genauer zu erfassen und dadurch auch zwischen den Lehrkräften weiter differenzieren zu können. Auch scheint es reizvoll, die Bereiche des erweiterten Lehrer-Handlungskompetenz-Modells zur Erklärung der Vorbereitungsqualität einzubeziehen, die in den Studien dieser Arbeit keinen Eingang in die Ergebnisse finden konnten. Dies betrifft vor allem die Domäne Wissen/Können, aber auch diejenigen gegenstandsbezogenen Überzeugungen, für die noch bessere Messinstrumente entwickelt werden müssen als die in den Studien A & B genutzten. Und schließlich steht im Raum dieser Perspektive auch die Frage, in welcher Form es zu Absprachen und Kooperationen im Vorfeld der Vorbereitungsphase kommt. Erfahrungen aus der eigenen Schulpraxis der letzten Jahre lassen den Autor vermuten, dass sich gerade in Bezug auf die Vorbereitungsphase ein großes Potenzial der Kooperation zeigt, dass gewinnbringend im Sinne von professionellen Lerngemeinschaften genutzt werden könnte. Auch dieser Bereich ist damit wie auch alle anderen Bereiche ein Beispiel für die Rolle des Testcoachings: Testcoaching im Rahmen von VERA8 ist nicht vorgesehen, aber seine Existenz ist ein bedeutender Indikator für notwendige Korrekturen am Instrument zentrale Vergleichsarbeiten.

Literaturverzeichnis

Abrams, L. M. & Madaus, G. F. (2003). The lessons of high-stakes testing. *Educational Leadership, 61* (3), 31–35.

Ackeren, I. v. (2003). *Nutzung großflächiger Tests für die Schulentwicklung. Exemplarische Analysen der Erfahrungen aus England, Frankreich und den Niederlanden* (Bildungsreform, Bd. 3). Berlin: Bundesministerium für Bildung und Forschung.

Ackeren, I. v. (2005). Vom Daten- und Informationsreichtum? - Erfahrungen mit standardisierten Vergleichstests in ausgewählten Nachbarländern. *Pädagogik* (5), 24–28.

Ackeren, I. v. (2007). Zentrale Abschlussprüfungen. Entstehung, Struktur und Steuerungsperspektive. *Pädagogik* (3), 12–15.

Ackeren, I. v. & Bellenberg, G. (2004). Parallelarbeiten, Vergleichsarbeiten und Zentrale Abschlussprüfungen. Bestandsaufnahme und Perspektiven. In H. G. Holtappels, K. Klemm, H. Pfeiffer, H. G. Rolff & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung* (Bd. 13, S. 125–159). Weinheim und München: Juventa Verlag.

Ackeren, I. v., Block, R., Klein, E. D. & Kühn, S. M. (2012). The impact of statewide exit exams: A descriptive case study of three German states with differing low stakes exam regimes. *Education Policy Analysis Archives, 20* (8), 1-30.

Ackeren, I. v., Heinrich, M. & Thiel, F (2013). *Evidenzbasiert Steuerung im Bildungssystem. Befunde aus dem BMBF-Förderschwerpunkt Steuerung im Bildungssystem (SteBis).* Die Deutsche Schule. 12. Beiheft. Münster: Waxmann.

Allalouf, A. & Ben-Shakhar, G. (1998). The effect of coaching the predictive validity of Scholastic Aptitude Tests. *Journal of Educational Measurement, 35* (1), 31–47.

Altrichter, H. (2000). Schulentwicklung und Professionalität. Bildungspolitische Entwicklungen und neue Anforderungen an Lehrer/innen. In J. Bastian, W. Helsper, S. Reh & C. Schelle (Hrsg.), *Professionalisierung im Lehrerberuf* (S. 147–163). Opladen: Leske + Budrich.

Altrichter, H. (2009). Datenfeedback und Unterrichtsentwicklung. Probleme eines Kernelements im "neuen Steuerungsmodell" für das Schulwesen. In W. Böttcher, J. N. Dicke & H. Ziegler (Hrsg.), *Evidenzbasierte Bildung. Wirkungsevaluation in Bildungspolitik und pädagogischer Praxis* (S. 211–226). Münster: Waxmann.

Altrichter, H. (2010). Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (Educational Governance, Bd. 7, S. 219–254). Wiesbaden: VS Verlag für Sozialwissenschaften.

Altrichter, H. & Heinrich, M. (2006). Evaluation als Steuerungsinstrument im Rahmen eines "neuen Steuerungsmodells" im Schulwesen. In W. Böttcher, H. G. Holtappels & M. Brohm

(Hrsg.), *Evaluation im Bildungswesen. Eine Einführung in Grundlagen und Praxisbeispiele* (Grundlagentexte Pädagogik, S. 51–64). Weinheim: Juventa Verlag.

Amrein, A. L. & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10* (18).

Amrein, A. L. & Berliner, D. C. (2003). The effects of high-stake testing on student motivation and learning. *Educational leadership,* 32–38.

Anderson, J. A. (2005). *Accountability in education*. Paris, Brüssel: UNESCO.

Aronson, E., Wilson, T. D. & Akert, R. M. (2004). Soziale Perzeption: Wie können wir andere Menschen verstehen? In E. Aronson, T. D. Wilson & R. M. Akert (Hrsg.), *Sozialpsychologie* (4., aktualisierte Aufl.). München: Pearson Studium.

Asendorpf, J. B. (2007). *Psychologie der Persönlichkeit*. Berlin, Heidelberg: Springer Medizin Verlag.

Ashby, J. & Sainsbury, M. (2001). *How do schools use national curriculum test results? A survey of the use of national curriculum test results in the management and planning of the curriculum at key stages 1 and 2:* National Foundation of Educational Research.

Au, W. (2007). High-stake testing and curricular control. A qualitative metasynthesis. *Educational Researcher, 36* (5), 258–267.

Baker, E. & Linn, R. L. (2004). Validity issues for accountability systems. In S. Fuhrman & R. F. Elmore (Hrsg.), *Redesigning accountability systems for education* (Critical issues in educational leadership series, S. 44–72). New York, NY: Teachers College Press.

Bangert-Drowns, R. L., Kulik, J. A. & Kulik, C. L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research, 85* (2), 89–99.

Barrick, M. R. & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44* (1), 1–26.

Bauer, K. O. (2009). Professionelles Selbst, Evaluieren und Innovieren. In T. Bohl & H. Kiper (Hrsg.), *Lernen aus Evaluationsergebnissen. Verbesserungen planen und implementieren* (S. 220–238). Bad Heilbrunn: Klinkhardt.

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U. et al. (2008). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV). Dokumentation der Erhebungsinstrumente*. Berlin: Max-Planck-Institut für Bildungsforschung.

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9* (4), 469–520.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A. et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47* (1), 133–180.

Baumert, J., Kunter, M., Brunner, M., Krauss, S., Blum, W. & Neubrand, M. (2004). Mathematikunterricht aus Sicht der PISA-Schülerinnen und -Schüler und ihrer Lehrkräfte. In PISA-Konsortium (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 314–354). Münster: Waxmann.

Baumert, J. & Watermann, R. (2000). Standardisierung durch Abiturprüfungen. Zentralabitur oder dezentrale Prüfungsorganisation? In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *TIMSSIII Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit* (TIMSS/III, Bd. 2, S. 341–351). Opladen: Leske + Budrich.

Baydar, N. (1990). *Effects of coaching on the validity of SAT: a simulation study.* Princeton, NJ: Educational Testing Service.

Becker, B. J. (1990). Coaching for Scholastic Aptitude Test: Further synthesis and appraisal. *Review of educational research, 60* (3), 373–417.

Bennewitz, H. (2009). Endlich nicht mehr Sisyphus. Ein positives berufliches Selbstkonzept gewinnen. *Pädagogik, 61* (10), 22–25.

Bereiter, C. & Scardamalia, M. (1993). *Surpassing ourselves. An inquiry into the nature and implications of expertise*. Chicago: Open Court.

Berkemeyer, N. (2010). *Die Steuerung des Schulsystems. Theoretische und praktische Explorationen*. Wiesbaden: VS, Verlag für Sozialwissenschaften.

Berkemeyer, N. & Bos, W. (2009). Professionalisierung im Spannungsfeld externer und interner Evaluation. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 529–541). Weinheim: Beltz.

Berkemeyer, N. & Müller, S. (2010). Schulinterne Evaluation – nur ein Instrument zur Selbststeuerung von Schulen? In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (Educational Governance, Bd. 7, S. 195–218). Wiesbaden: VS Verlag für Sozialwissenschaften

Berliner, D. C. (1990). Whats all the fuss about instructional time? In M. Ben-Peretz & R. Bromme (Hrsg.), *The Nature of time in schools. Theoretical concepts, practitioner perceptions* (S. 3–35). New York: Teachers College Press.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35,* 463–482.

Besser, M. & Krauss, S. (2009). Zur Professionalität der Expertise. In O. Zlatkin-Troitschanskaia, K. Beck, D. Ser, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung .* Weinheim: Beltz.

Block, R., Klein, E. D., Ackeren, I. van & Kühn, S. M. (2011). Leistungseffekte des Zentralabiturs? Eine kritische Auseinandersetzung mit bildungsökonomischen Interpretationen zu den Effekten der Prüfungsorganisation auf der Basis von PISA E 2003-Daten. *bildungsforschung, 8* (1), 215–238.

Blömeke, S., Kaiser, G., Döhrmann, M. & Lehmann, R. (2010). Mathematisches und mathematikdidaktisches Wissen angehender Sekundarstufen-I-Lehrkräfte im internationalen Vergleich. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* (S. 197–238). Münster: Waxmann.

Blömeke, S., Kaiser, G. & Lehmann, R. (Hrsg.). (2008). *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und –refendare. Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung*. Münster: Waxmann.

Blömeke, S., Kaiser, G. & Lehmann, R. (Hrsg.). (2010a). *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich*. Münster: Waxmann.

Blömeke, S., Kaiser, G. & Lehmann, R. (2010b). TEDS-M 2008 Sekundarstufe I: Ziele, Untersuchungsanlage und zentrale Ergebnisse. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* (S. 11–38). Münster: Waxmann.

Blömeke, S. & König, J. (2010). Messung des pädagogischen Wissens: Theoretischer Rahmen und Teststruktur. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* (S. 239–263). Münster: Waxmann.

Blömeke, S., Müller, C., Felbrich, A. & Kaiser, G. (2008). Epistemologische Überzeugungen zur Mathematik. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und –refendare. Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung* . Münster: Waxmann.

Blum, W. (2010). *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen ; mit CD-ROM* (4. Aufl.). Berlin: Cornelsen Scriptor.

Blum, W., Krauss, S. & Neubrand, M. (2008, März). *Zusammenhänge des Professionswissens mit Lehrermerkmalen, Unterrichtsqualität und Leistungszuwächsen der Schüler,* Budapest.

Böhm-Kasper, O. (2004). *Schulische Belastung und Beanspruchung. Eine Untersuchung von Schülern und Lehrern am Gymnasium* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 43). Münster: Münster: Waxmann.

Böhm-Kasper, O., Bos, W., Jaeckel, S. & Weishaupt, H. (2000). *Skalenhandbuch zur Belastung von Schülern und Lehrern. Das Erfurter Belastungs-Inventar (EBI)* (Erfurter Materialien und Berichte zur Entwicklung des Bildungswesens). Erfurt: Pädagogische Hochschule.

Böhm-Kasper, O., Bos, W., Körner, S. C. & Weishaupt, H. (2001). *Sind 12 Schuljahre stressiger? Belastung und Beanspruchung von Lehrern und Schülern am Gymnasium* (W. Bos, Hrsg., Veröffentlichungen der Max-Traeger-Stiftung, Bd. 35). Weinheim: Juventa Verlag

Bond, L. (1993). The effects of special preparation on measures of scholastic ability. In R. L. Linn (Hrsg.), *Educational measurement* (3. ed., S. 429–444). New York, NY: American Council on Education [u.a.].

Bond, L. & Harman, A. (1995). Test-taking strategies: the effects of coaching and practice on measures of intelligence. In R. J. Sternberg (Hrsg.), *Encyclopedia of Human Intelligence* (5. Aufl., S. 1073–1077). New York: Macmillan [u.a.].

Bong, M. & Clark, R. E. (1999). Comparison between self-concept und self-efficacy in academic motivation research. *Educational Psychologist, 34* (3), 139–153.

Bönsch, M. (2008). Faktoren für die Lehrergesundheit. *Schulmagazin 5 bis 10* (3), 53–56.

Bonsen, M., Büchter, A. & Peek, R. (2006). Datengestützte Schul- und Unterrichtsentwicklung - Bewertungen der Lernstandserhebungen in NRW durch Lehrerinnen und Lehrer. In W. Bos, H. G. Holtappels, H. Pfeiffer, H. G. Rolff & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung* (Bd. 14, S. 125–148). Weinheim und München: Juventa Verlag.

Bonsen, M. & Gathen, J. v. (2004). Schulentwicklung und Testdaten – die innerschulische Verarbeitung von Leistungsrückmeldungen. In H. G. Holtappels, K. Klemm, H. Pfeiffer, H. G. Rolff & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung* (Bd. 13, S. 225–252). Weinheim und München: Juventa Verlag.

Bos, W. (1989). Reliabilität und Validität in der Inhaltsanalyse – ein Beispiel zur Kategorienoptimierung in der Analyse chinesischer Textbücher für den muttersprachlichen Unterricht von Auslandschinesen. In W. Bos & C. Tarnai (Hrsg.), *Angewandte Inhaltsanalyse in empirischer Pädagogik und Psychologie* (Waxmann Wissenschaft, S. 62–72). Münster: Waxmann.

Böttcher, W. (2006). Bildungsstandards und Evaluation im Paradigma der Outputsteuerung. In W. Böttcher, H. G. Holtappels & M. Brohm (Hrsg.), *Evaluation im Bildungswesen. Eine Einführung in Grundlagen und Praxisbeispiele* (Grundlagentexte Pädagogik, S. 39–49). Weinheim: Juventa Verlag.

Böttger-Beer, M. & Koch, E. (2008). Externe Schulevaluation in Sachsen – ein Dialog zwischen Wissenschaft und Praxis. In W. Böttcher, W. Bos, H. Döbert & H. G. Holtappels (Hrsg.), *Bildungsmonitoring und Bildungscontrolling in nationaler und internationaler Perspektive. Dokumentation zur Herbsttagung der Kommission Bildungsorganisation, -planung, -recht (KBBB)* (S. 253–264). Münster: Waxmann.

Bromme, R. (1997). Kompetenzen, Funktionen und unterrichtliches Handeln des Lehrers. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule* (Enzyklopädie der Psychologie Praxisgebiete Pädagogische Psychologie, S. 177–212). Göttingen: Hogrefe.

Bromme, R. (2008). Lehrerexpertise. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie* (S. 159–167). Göttingen: Hogrefe.

Bromme, R. & Haag, L. (2008). Forschung zur Lehrerpersönlichkeit. In W. Helsper & J. Böhme (Hrsg.), *Handbuch der Schulforschung* (2., durchgesehene und erweiterte Auflage., S. 803–819). Wiesbaden: VS Verlag.

Bromme, R. & Rambow, R. (2001). Experten-Laien-Kommunikation als Gegenstand der Expertiseforschung: Für eine Erweiterung des psychologischen Bildes vom Experten. In R. K. Silbereisen & M. Reitzle (Hrsg.), *Psychologie 2000. Bericht über den 42. Kongress der Deutschen Gesellschaft für Psychologie in Jena 2000* (S. 541–550). Lengerich: Pabst Science Publishers.

Bromme, R. & Rheinberg, F. (2006). Lehrende in Schulen. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie. Ein Lehrbuch* (5., vollst. überarb. Aufl., S. 296–332). Weinheim: Beltz PVU.

Brophy, J. (1999). *Teaching* (Educational practices series, Bd. 1). Brüssel, Genf: International Academy of Education, International Bureau of Education.

Brown, G. T. L. (2011). Teachers' conception of assessment: Comparing primary & secondary teacher in New Zealand. *Assessment Matters, 3, 45-70*.

Brown, G. T. L. & Harris, L. R. (2009). Unintended consequences of using tests to improve learning: How improvement-oriented resources heighten conceptions of assessment as school accountability. *Journal of MultiDisciplinary Evaluation, 6* (12), 68–91.

Brunner, M., Artelt, C., Krauss, S. & Baumert, J. (2007). Coaching for the PISA test. *Learning and instruction, 17,* 111–122.

Brunner, M., Kunter, M., Krauss, S., Baumert, J., Blum, W., Dubberke, T. et al. (2006). Welche Zusammenhänge bestehen zwischen dem fachspezifischen Professionswissen von Mathematiklehrkräften und ihrer Ausbildung sowie beruflichen Fortbildungen?. *Zeitschrift für Erziehungswissenschaft, 9* (4), 521–544.

Brunner, M., Kunter, M., Krauss, S., Klusmann, U., Baumert, J., Blum, W. et al. (2006). Die professionelle Kompetenz von Mathematiklehrkräften. Konzeptualisierung, Erfassung und Bedeutung für den Unterricht. Eine Zwischenbilanz des COACTIV-Projekts. In M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* . Münster: Waxmann.

Büchter, A. & Leuders, T. (2005). Zentrale Tests und Unterrichtsentwicklung …bei guten Aufgaben und gehaltvollen Rückmeldungen keine Widerspruch. *Pädagogik, 57* (5), 14–18.

Büchter, A. & Leuders, T. (2007). *Mathematikaufgaben selbst entwickeln. Lernen fördern - Leistung überprüfen* (3. Aufl.). Berlin: Cornelsen Scriptor.

Bühner, M. (2010). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson Studium.

Bunting, B. P. & Mooney, E. (2001). The effects of practice and coaching on test results for educational selection at eleven years of age. *Educational Psychology, 21* (3), 243–253.

Burkard, C. & Peek, R. (2004). Anforderungen an zentrale Lernstandserhebungen – ein Werkstattbericht aus Nordrhein-Westfalen. *Pädagogik, 6,* 24–27.

Burns, M. K., Courtad, C. A., Hoffman, H. & Folger, W. (2004). A comparison of district-level variables and state accountability test results for public elementary and middle schools. *Psychology and Education, 41* (2), 17-16.

Carnoy, M. (2005). Have state accountability and high-stake tests influenced student progressing. *Educational measurement* (4), 19–31.

Carnoy, M. & Loeb, S. (2004). Does external accountability affect students outcome? A cross-state analysis. In S. Fuhrman & R. F. Elmore (Hrsg.), *Redesigning accountability systems for education* (Critical issues in educational leadership series, S. 189–219). New York, NY: Teachers College Press.

Carstensen, C. H., Knoll, S., Rost, J. & Prenzel, M. (2004). Technische Grundlagen. In PISA-Konsortium (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 371–387). Münster: Waxmann.

Cheng, L. & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. J. Watanabe & A. Curtis (Hrsg.), *Washback in language testing. Research contexts and methods* (S. 3–17). Mahwah, NJ: Erlbaum.

Clausen, M. (2002). *Unterrichtsqualität: eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 29). Münster: Waxmann.

Clausen, M. (2007). *Einzelschulwahl. Zur Wahl der konkreten weiterführenden Einzelschule aus der Sicht von Bildungsnachfragenden und Bildungsanbietenden*. Habilitationsschrift, Universität Mannheim. Mannheim

Coe, R. (1998). Can feedback improve teaching? A review of the social science literature with a view to identifying the conditions under which giving feedback to teachers will result in improved performance. *Research Papers in Education, 13* (1), 43–66.

Corbalan, G., Kester, L. & Merrienboer, J. J. G. van. (2009). Dynamic task selection: Effects of feedback and learner control on efficiency and motivation. *Learning and instruction, 19,* 455–465.

Costa, P. T. & McCrae, R. R. (1992). *Revised NEO Personality Inventory and NEO Five Factor Inventory professional manual*. Odessa: Psychological Assessment Resources.

Crundwell, R. M. (2005). Alternative strategies for large scale student assessment in Canada: Is value-added assessment one possible answer. *Canadian Journal of Educational Administration and Policy, 41.* Zugriff am 17.10.2011. Verfügbar unter http://umanitoba.ca/publications/cjeap/articles/crundwell.html.

Dann, H. D. (2008). Lehrerkognitionen und Handlungsentscheidungen. In M. K. W. Schweer (Hrsg.), *Lehrer-Schüler-Interaktion. Inhaltsfelder, Forschungsperspektiven und methodische Zugänge* (Schule und Gesellschaft, Bd. 24, 2., vollständig überarbeitete Auflage., S. 177–205). Wiesbaden: VS Verlag.

Darling-Hammond, L. (2004). Standards, accountability and school reform. *Teachers College record, 106* (6), 1047–1085.

Dauenheimer, D., Stahlberg, D., Frey, D. & Petersen, L. E. (2002). Die Theorie des Selbstwertschutzes und der Selbstwerterhöhung. In D. Frey & M. Irle (Hrsg.), *Motivations-, Selbst- und Informationsverarbeitungstheorien* (Psychologie-Lehrtexte, Bd. 3, 2., vollst. überarb. und erw. Aufl., S. 159–190). Bern: Huber.

Davier, M. von. (1997). Bootstrapping Goodness-of-Fit statistics for sparse categorical data - results of a Monte-Carlo-Study. *Methods of Psychological Research Online, 2* (2), 29–48.

Dedering, K. (2011). Hat Feedback eine positive Wirkung? Zur Verarbeitung extern erhobener Leistungsdaten in Schulen. *Unterrichtswissenschaft, 39* (1), 63–83.

Demerouti, E., Bakker, A. B., Nachreimer, F. & Schaufeli, W. B. (2001). The job demands-resources model of burnout. *Journal of Applied Psychology, 86,* 499–512.

DerSimonian, R. & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: a meta-analysis. *Harvard Educational Review, 52* (1), 1–15.

Dickhäuser, O. (2006). Editorial zum Themenschwerpunkt Fähigkeitsselbstkonzepte. Entstehung, Auswirkung, Förderung. *Zeitschrift für Pädagogische Psychologie, 20* (1/2), 5–8.

Dick, R. van. (2006). *Stress und Arbeitszufriedenheit bei Lehrerinnen und Lehrern. Zwischen "Horrorjob" und Erfüllung* (2., leicht veränd. Aufl.). Marburg: Tectum-Verlag.

Diedrich, M., Thußbas, C. & Klieme, E. (2002). Professionelles Lehrerwissen und selbstberichtete Unterrichtspraxis im Fach Mathematik. In M. Prenzel & J. Doll (Hrsg.), *Bildungsqualität von Schule. Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen.* (Zeitschrift für Pädagogik, Beiheft 45, S. 107–123). Weinheim: Beltz.

Diekmann, A. & Jann, B. (2000). *Anreizformen und Rücklaufquoten bei postalischen Befragungen. Eine Prüfung der Reziprozitätshypothese.* Bern: Institut für Soziologie der Universität Bern. Zugriff am 27.03.2011. Verfügbar unter http://www.socio.ethz.ch/people/andreasd/working_papers/TELEFO2.pdf.

Diemer, T. & Kuper, H. (2011). Formen innerschulischer Steuerung mittels zentraler Lernstandserhebungen. *Zeitschrift für Pädagogik, 56* (4), 554–571.

Ditton, H. (2000). Qualitätskontrolle und Qualitätssicherung in Schule und Unterricht. In A. Helmke, W. Hornstein & E. Terhart (Hrsg.), *Qualität und Qualitätssicherung im Bildungsbereich: Schule, Sozialpädagogik, Hochschule* (Zeitschrift für Pädagogik Beiheft, Bd. 41, S. 73–92). Weinheim: Beltz.

Ditton, H. (2009). Unterrichtsqualität. In K. H. Arnold, U. Sandfuchs & J. Wiechmann (Hrsg.), *Handbuch Unterricht* (2. aktualisierte Aufl., S. 177–183). Stuttgart: UTB.

Ditton, H. & Arnoldt, B. (2004). Schülerbefragung zum Fachunterricht - Feedback an Lehrkräfte. *Empirische Pädagogik, 18* (1), 115–139.

Ditton, H. & Arnoldt, B. (2004a). Wirksamkeit von Schülerfeedback zum Fachunterricht. In J. Doll & M. Prenzel (Hrsg.), *Bildungsqualität von Schule. Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung* (S. 152–170). Münster: Waxmann.

Dobbelstein, P. & Peek R. (2008). Lernstandserhebungen und zentrale Prüfungen im Kontext der Qualitätsanalyse. In S. Müller, D. Dedering & W. Bos (Hrsg.), *Schulische Qualitätsanalyse in Nordrhein-Westfalen. Konzepte, erste Erfahrungen, Perspektive.* Köln: LinkLuchterhand.

Döhrmann, M., Kaiser, G. & Blömeke, S. (2010). Messung des mathematischen und mathematikdidaktischen Wissens: Theoretischer Rahmen und Teststruktur. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* (S. 169–196). Münster: Waxmann.

Dubberke, T., Kunter, M., McElvany, N., Brunner, M. & Baumert., J. (2008). Lerntheoretische Überzeugungen von Mathematiklehrkräften. Einflüsse auf die Unterrichtsgestaltung und den Lernerfolg von Schülerinnen und Schülern. *Zeitschrift für Pädagogische Psychologie, 22* (3-4), 193–206.

Dubs, R. (2006). Qualitätsmanagement. In H. Buchen & H.-G. Rolff (Hrsg.), *Professionswissen Schulleitung .* Weinheim: Beltz.

Edelstein, W. (1998). Selbstwirksamkeit im Kontext der Schulreform. *Pädagogische Führung, 9* (2), 56–59.

Edelstein, W. (2002). Selbstwirksamkeit, Innovation und Schulreform. Zur Diagnose der Situation. In M. Jerusalem & D. Hopf (Hrsg.), *Selbstwirksamkeit und Motivationsprozesse in Bildungsinstitutionen.* (Zeitschrift für Pädagogik, Beiheft 44, S. 13–27). Weinheim: Beltz.

Erickson, G. & Lander, R. (2007). Der Kitt, der ein wachsendes System zusammenhält? Nationale Tests als Kern der Qualitätssicherung in Schweden. *Pädagogik* (3), 32–35.

Ericsson, K. A. & Charness, N. (1994). Expert performance. Its structure and acquisition. *American Psychologist, 49,* 725–747.

Ericsson, K. A., Krampe, R. & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363–406.

Fend, H. (2008). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Fend, H. (2009). *Neue Theorie der Schule. Einführung in das Verstehen von Bildungssystemen* (2., durchgesehene Auflage.). Wiesbaden: VS Verlag für.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations* (7), 117–140.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Fischer, P., Frey, D., Peus, C. & Kastenmueller, A. (2008). The theory of cognitive dissonance: State of the science and directions for future research. In P. Meusburger, M. Welker & E. Wunder (Hrsg.), *Clashes of Knowledge. Orthodoxies and Heterodoxies in Science and Religion* (Knowledge and Space, Bd. 1, S. 189–198). Dordrecht: Springer Science + Business Media B.V.

Flippo, R. F., Becker, M. J. & Wark, D. M. (2000). Preparing for and taking Tests. In R. F. Flippo & D. C. Caverly (Hrsg.), *Handbook of college reading and study strategy research* (S. 211–260). Mahwah, NJ: Lawrence Erlbaum Associates.

Frese, M. & Zapf, D. (1994). Action as the core of work psychology: A German Approach. In H. C. Triandis & M. D. L. M. H. Dunnette (Hrsg.), *Handbook of industrial and organizational psychology. Volume 4* (2. ed., S. 271–339). Palo Alto, Calif.: Consulting Psychologists Press.

Fried, L. (2002). *Pädagogisches Professionswissen und Schulentwicklung. Eine systemtheoretische Einführung in Grundkategorien der Schultheorie* (Basistexte Erziehungswissenschaft. Weinheim: Juventa Verlag.

Fuchs, T. & Wößmann, L. (2007). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics, 32* (2-3), 433–464.

Fuhrman, S. (2004). Introduction of educational standards in German-speaking countries. In S. Fuhrman & R. F. Elmore (Hrsg.), *Redesigning accountability systems for education* (Critical issues in educational leadership series, S. 3–13). New York, NY: Teachers College Press.

Fussangel, k., Ditzinger, V., Böhm-Kasper, O. & Gräsel, C. (2010). Kooperation, Belastung und Beanspruchung von Lehrkräften an Halb- und Ganztagsschulen. *Unterrichtswissenschaft, 38* (1), 51–67.

Garner, R. (2005). Post-It® Note persuasion: A sticky influence. *Journal of Consumer Psychology, 15* (3), 230-237.

Gathen, J. v. d. (2011). *Leistungsrückmeldungen bei Large-Scale-Assessments und Vollerhebungen. Rezeption und Nutzung am Beispiel von DESI und lernstand* (Internationale Hochschulschriften, Bd. 552). Münster: Waxmann.

Glaser, R. (1996). Changing the agency for learning: Acquiring expert performance. In K. A. Ericsson (Hrsg.), *The road to excellence. The aquisition of expert performance in the arts and sciences, sports, and games* (S. 303–311). Mahwah NJ: Erlbaum.

Gollwitzer, M., Mossbrugger, H. & Kelave, A. (2008). Latent-Class-Analysis. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* . Heidelberg: Springer.

Gräsel, C., Fußangel, K. & Pröbstel, C. (2006). Lehrkräfte zur Kooperation anregen – eine Aufgabe für Sisyphos? *Zeitschrift für Pädagogik* (2), 205–219.

Green, A. (2006). Watching for washback: Observing the influence of the international English language testing system academic writing test in the classroom. *Language Assessment Quarterly, 3* (4), 333–368.

Greve, A. (2011, 17. Juni). *Testmodelle bei VERA 8* (E-Mail).

Groß Ophoff, J. (2013). *Lernstandserhebungen: Reflexion und Nutzung* (Bd. 85). Münster: Waxmann.

Groß Ophoff, J., Hosenfeld, I. & Koch, U. (2007). Formen der Ergebnisrezeption und damit verbundene Schul- und Unterrichtsentwicklung. *Empirische Pädagogik, 21* (4), 411–428.

Haag, L. (2004). Lehrerkognitionsforschung - notwendige, aber nicht hinreichende Bedingungen einer professionellen Lehrerbildung. In M. Eckert (Hrsg.), *Studien zur Dynamik des Berufsbildungssystems. Forschungsbeiträge zur Struktur-, Organisations- und Curriculumentwicklung* (Schriftenreihe der Sektion Berufs- und Wirtschaftspädagogik der DGfE, 1. Aufl., S. 159–171). Wiesbaden: VS Verlag für Sozialwissenschaften.

Haag, L. (2006). Nachhilfeunterricht. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (Schlüsselbegriffe, 3., überarb. und erw. Aufl.). Weinheim: Beltz PVU.

Hackman, J. R. & Oldham, G. R. (1980). *Work redesign* (Addison-Wesley series on organization development). Reading, Mass.: Addison-Wesley.

Hahn, J. (2008). *Testcoaching im Rahmen der Lernstandserhebungen. Hausarbeit im Rahmen des Ersten Staatsexamens für das Lehramt an Gymnasien und Gesamtschulen*, unveröffentlicht.

Hakanen, J., Bakker, A. & Schaufeli, W. B. (2006). Burnout and work engagement among teachers. *Journal of School Psycholog, 43,* 495–513.

Hanushek, E. A. & Raymond, M. E. (2004). The effect of school accountability systems on the level and distribution of student achievement. *Journal of the European Economic Association, 2* (2-3), 406–415.

Hanushek, E. A. & Raymond, M. E. (2006). School accountability and student performance. *Regional Economic Development, 2* (1), 51–61.

Harazd, B., Gieske, M. & Rolff, H.-G. (2009). *Gesundheitsmanagement in der Schule. Lehrergesundheit als neue Aufgabe der Schulleitung - eine Veröffentlichung der Dortmunder Akademie für Pädagogische Führungskräfte (DAPF) der Technischen Universität Dortmund* (Schule und Gesundheit). Köln: LinkLuchterhand.

Hartung-Beck, V. (2009). *Schulische Organisationsentwicklung und Professionalisierung - Folgen von Lernstandserhebungen an Gesamtschulen*. Wiesbaden: VS Verlag.

Hascher, T. & Krapp, A. (2009). Emotionale Voraussetzungen der Entwicklung der Professionalität von Lehrenden. In O. Zlatkin-Troitschanskaia, K. Beck, D. Ser, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 365–375). Weinheim: Beltz.

Hattie, J. (1992). *Self concept*. Hillsdale, NJ: Erlbaum.

Hattie, J. & Marsh, H. W. (1996). Future directions in Self-Concept research. In B. A. Bracken (Hrsg.), *Handbook of self-concept. Developmental, social, and clinical considerations*. New York: Wiley.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of educational research, 77* (1), 81-112.

Helmke, A. (2004). Von der Evaluation zur Innovation. Pädagogische Nutzbarmachung von Vergleichsarbeiten in der Grundschule. *Seminar* (2), 90–112.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (2., aktualisierte Aufl.). Seelze-Velber: Klett Kallmeyer.

Helmke, A. & Hosenfeld, I. (2003). Vergleichsarbeiten (VERA): eine Standortbestimmung zur Sicherung schulischer Kompetenz – ein viel beachteter Ansatz aus Rheinland-Pfalz. *SchulVerwaltung NRW, 4,* 107–110.

Helmke, A. & Hosenfeld, I. (2005). Standardbezogene Unterrichtsevaluation. In G. Brägger, B. Bucher & N. Landwehr (Hrsg.), *Schlüsselfragen zur externen Schulevaluation* (Pädagogik, S. 127–151). Bern: h.e.p. Verlag.

Hense, J. U. & Mandl, H. (2009). Evaluations- und Selbstevaluationskompetenz von Lehrenden. Warum benötigen Lehrende Evaluationskompetenz? In O. Zlatkin-Troitschanskaia, K. Beck, D. Ser, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 129–139). Weinheim: Beltz.

Herman, J. L. (2004). The effects of testing on instruction. In S. Fuhrman & R. F. Elmore (Hrsg.), *Redesigning accountability systems for education* (Critical issues in educational leadership series, S. 141–166). New York, NY: Teachers College Press.

Hertel, S. & Bruder, S. S. B. (2009). Beratungs- und Gesprächsführungskompetenz von Lehrkräften. In O. Zlatkin-Troitschanskaia, K. Beck, D. Ser, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 117–128). Weinheim: Beltz.

Heubert, J. P. (2004). High-stakes testing in a changing environment. Disparate impact, opportunity to learn, and current legal protections. In S. Fuhrman & R. F. Elmore (Hrsg.), *Redesigning accountability systems for education* (Critical issues in educational leadership series, S. 220–241). New York, NY: Teachers College Press.

Heymann, H. & Pallack, A. (2007). Aufgabenkonstruktion für die Lernstandserhebung Mathematik. In Ministerium für Schule und Weiterbildung (Hrsg.), *Lernstandserhebungen Mathematik in Nordrhein-Westfalen. Impulse zum Umgang mit zentralen Tests* (S. 14–46). Stuttgart, Leipzig: Klett Verlag.

Hobfoll, S. E. & Buchwald, P. (2004). Die Theorie der Ressourcenerhaltung und das multiaxiale Copingmodell - eine innovative Stresstheorie. In P. Buchwald, C. Schwarzer & S. E. Hobfoll (Hrsg.), *Stress gemeinsam bewältigen. Ressourcenmanagement und multiaxiales Coping* (S. 11–26). Göttingen: Hogrefe.

Hofmann, J. M. & Preiser, S. (1989). *Veränderungen von Kontrollüberzeugungen während des Lehramtspraktikums. Untersuchungen zum Einfluß der Praktikumssituation und der induzierten Reflexion über Handlungsbedingungen*. Frankfurt a. M.: Universität, Institut für Pädagogische Psychologie.

Hosenfeld, A. (2010). *Führt Unterrichtsrückmeldung zu Unterrichtsentwicklung? Die Wirkung von videographischer und schriftlicher Rückmeldung bei Lehrkräften der vierten Jahrgangsstufe* (Empirische Erziehungswissenschaft, Bd. 22). Münster: Waxmann.

Hossiep, R. & Paschen, M. (2003). *Das Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP)*. Göttingen: Hogrefe.

Huber, C., Späni, M., Schmellentin, C. & Criblez, L. (2006). *Bildungsstandards in Deutschland, Österreich, England, Australien, Neuseeland und Südostasien. Literaturbericht zu Entwicklung, Implementation und Gebrauch von Standards in nationalen Schulsystemen*. Aarau: Fachhochschule Nordwestschweiz. Zugriff am 07.10.2011. Verfügbar unter http://www.edudoc.ch/static/web/arbeiten/harmos/lit_analyse_1.pdf.

Hugener, I. (2008). *Inszenierungsmuster im Unterricht und Lernqualität. Sichtstrukturen schweizerischen und deutschen Mathematikunterrichts in ihrer Beziehung zu Schülerwahrnehmung und Lernleistung - eine Videoanalyse* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 68). Münster: Waxmann (Univ., Diss.--Zürich, 2007.).

Ilgen, D. R. & Davis, C. A. (2000). Bearing bad news: Reactions to negative performance feedback. *Applied Psychology, 49* (3), 550–565.

Ingram, D., Seashore Louis, K. & Schroeder, R. G. (2004). Accountability policies and teacher decision making. Barriers to the use of data to improve practice. *Teachers College record, 106* (6), 1258–1287.

Jäger, D. J. (2012). Herausforderung Zentralabitur: Unterrichtsinhalte variieren und an Prüfungsthemen anpassen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern .* Wiesbaden: VS Verlag für Sozialwissenschaften.

Jakobs, B. (2008). Was wissen wir über die Lernwirksamkeit von Aufgabenstellungen und Feedback. In J. Thonhauser (Hrsg.), *Aufgaben als Katalysatoren von Lernprozessen. Eine*

zentrale Komponente organisierten Lehrens und Lernens aus der Sicht von Lernforschung, allgemeiner Didaktik und Fachdidaktik (S. 99–114). Münster: Waxmann.

Jankisz, E. & Moosbrugger, H. (2008). Planung und Entwicklung von psychologischen Tests und Fragebogen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 27–72). Heidelberg: Springer.

Jensen, M. C. & Meckling, W. H. (1976). Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics, 3* (4), 305–360.

Jerusalem, M. (2002). Einleitung. In M. Jerusalem & D. Hopf (Hrsg.), *Selbstwirksamkeit und Motivationsprozesse in Bildungsinstitutionen.* (Zeitschrift für Pädagogik, Beiheft 44, S. 8–12). Weinheim: Beltz.

Jerusalem, M. & Schwarzer, R. (1992). Self-efficacy as a resource factot in stress appraisal processes. In R. Schwarzer (Hrsg.), *Self-efficacy. thought control of action* (S. 195–213). Washington: Hemispehre Publ. Corp.

Johnson, R. (1998). Toward a theoretical model of evaluation utilization. *Evaluation and Program Planning, 21* (1), 93–110.

Jones, B. D. (2007). The unintended outcome of high-stakes testing. *Journal of Applied School Psychology, 23* (2), 65–86.

Jonge, J. de, Linden, S. van der, Schaufeli, W., Peter, R. & Siegrist, J. (2008). Factorial invariance and stability of the Effort-Reward Imbalance Scales: A longitudinal analysis of two samples with different time lags. *International Journal of Behavioral Medicine, 15,* 62–72.

Jürges, H. & Schneider, K. (2008). Ressourcen und Anreize im Bildungswesen. Aufgaben und Handlungsmöglichkeiten des Staates aus Sicht der Bildungsökonomie. *Zeitschrift für Erziehungswissenschaft* (11), 234–252.

Keller-Schneider, M. (2010). *Entwicklungsaufgaben im Berufseinstieg von Lehrpersonen. Beanspruchung durch berufliche Herausforderungen im Zusammenhang mit Kontext- und Persönlichkeitsmerkmalen* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 78). Berlin: Waxmann.

Kempfert, G. & Rolff, H. G. (2005). *Qualität und Evaluation - ein Leitfaden für pädagogisches Qualitätsmanagement*. Weinheim und Basel: Beltz Verlag.

Kiel, E., Geider, F. J. & Jünger, W. (2004). Motivation, Selbstkonzepte und Lehrerberuf. Studienwahl und Berufsperspektiven bei Studierenden für das Lehramt an Grund-, Haupt-, und Realschulen. *Die Deutsche Schule, 96* (2), 223–233.

Kieser, A. & Ebers, M. (2006). *Organisationstheorien* (6. Aufl.). Stuttgart: Kohlhammer.

Kiper, H. (2009). Schulentwicklung im Rahmen von Kontextsteuerung - Welche Hinweise geben (durch Evaluation und Vergleichsarbeiten gewonnene) Daten für ihre Ausrichtung?

In T. Bohl & H. Kiper (Hrsg.), *Lernen aus Evaluationsergebnissen. Verbesserungen planen und implementieren* (S. 13–28). Bad Heilbrunn: Klinkhardt.

Klein, E.D. (2013): Statewide Exit Exams, Governance, and School Development. An International Comparison. Münster: Waxmann.

Klein, E. D., Kühn, S. M., Ackeren, I. v. & Block, R. (2009). Wie zentral sind zentrale Prüfungen? Abschlussprüfungen am Ende der Sekundarstufe II im nationalen und internationalen Vergleich. *Zeitschrift für Pädagogik, 55* (4), 596–621.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise* (Bundesministerium für Bildung und Forschung, Hrsg.), Bonn

Klieme, E. & Reusser, K. (2003). Unterrichtsqualität und mathematisches Verständnis im internationalen Vergleich - Ein Forschungsprojekt und erste Schritte zur Realisierung. *Unterrichtswissenschaft, 31* (3), 194–205.

Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary Feedback Intervention Theory. *Psychological Bulletin, 119* (2), 254-284.

Klusmann, U. (2008). Gesamtdiskussion. In U. Klusmann (Hrsg.), *Berufliches Beanspruchungserleben und Unterrichtsverhalten von Lehrkräften. Zur Rolle persönlicher und institutioneller Ressourcen .* Berlin: Max-Planck-Institut für Bildungsforschung.

Klusmann, U. (2008a). Theoretischer Rahmen der Arbeit. In U. Klusmann (Hrsg.), *Berufliches Beanspruchungserleben und Unterrichtsverhalten von Lehrkräften. Zur Rolle persönlicher und institutioneller Ressourcen .* Berlin: Max-Planck-Institut für Bildungsforschung.

Klusmann, U., Kunter, M. & Trautwein, U. (2009). Die Entwicklung des Beanspruchungserlebens bei Lehrerinnen und Lehrern in Abhängigkeit beruflicher Verhaltensstile. *Psychologie in Erziehung und Unterricht* (3), 200–212.

Klusmann, U., Kunter, M., Trautwein, U. & Baumert, J. (2006). Lehrerbelastung und Unterrichtsqualität aus der Perspektive von Lehrenden und Lernenden. *Zeitschrift für Pädagogische Psychologie,* 161–173.

Koch, U. (2011). *Verstehen Lehrkräfte Rückmeldungen aus Vergleichsarbeiten? Datenkompetenz von Lehrkräften und die Nutzung von Ergebnisrückmeldungen aus Vergleichsarbeiten* (Empirische Erziehungswissenschaft 31). Münster: Waxmann.

Koch, U., Groß Ophoff, J., Hosenfeld, I. & Helmke, A. (2006). Qualitätssicherung: Von der Evaluation zur Schul- und Unterrichtsentwicklung – Ergebnisse der Lehrerbefragung zur Auseinandersetzung mit der VERA-Rückmeldung. In F. Eder, A. Gastager & F. Hofmann (Hrsg.), *Qualität durch Standards? Beiträge zum Schwerpunktthema der 67. Tagung der AEPF.* (S. 187–199). Münster: Waxmann.

Kohler, B. (2009). Umgang von Lehrer/innen, Eltern und Schulaufsicht mit den Ergebnissen internationaler Schulleistungsmessungsstudien. In T. Bohl & H. Kiper (Hrsg.), *Lernen aus*

Evaluationsergebnissen. Verbesserungen planen und implementieren (S. 81–96). Bad Heilbrunn: Klinkhardt.

Kohler, B. & Schrader, W. (2004). Ergebnisrückmeldung und Rezeption. Von der externen Evaluation zur Entwicklung von Schule und Unterricht. *Empirische Pädagogik, 18* (1), 3–17.

Köller, O. (2008). *Erläuterungen zur Bereitstellung eines normierten Aufgabenpools für kompetenzbasierte Lernstandserhebungen im Fach Mathematik in der 8. Jahrgangsstufe.* Zugriff am 17.06.2008. Verfügbar unter http://www.isq-bb.de/pdf/vera8/Testkonzeption_Lernstand_M8.pdf.

Köller, O. & Möller, J. (2006). Selbstwirksamkeit. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (Schlüsselbegriffe, 3., überarb. und erw. Aufl., S. 767–774). Weinheim: Beltz PVU.

Koretz, D. (2008). Test-based educational accountability. *Zeitschrift für Pädagogik* (6), 777–790.

Krapp, A. & Ryan, R. M. (2002). Selbstwirksamkeit und Lernmotivation. Eine kritische Betrachtung der Theorie von Bandura aus Sicht der Selbstbestimmungstheorie und der pädagogisch-psychologischen Interessentheorie. In M. Jerusalem & D. Hopf (Hrsg.), *Selbstwirksamkeit und Motivationsprozesse in Bildungsinstitutionen.* (Zeitschrift für Pädagogik, Beiheft 44). Weinheim: Beltz.

Krause, U. M., Stark, R. & Mandl, H. (2004). Förderung des computerbasierten Wissenserwerbs durch kooperatives Lernen und eine Feedbackmaßnahme. *Zeitschrift für Pädagogische Psychologie, 18* (2), 125–136.

Krauss, S., Kunter, M., Brunner, M., Baumert, J., Blum, W., Neubrand, M. et al. (2004). COACTIV: Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz. In J. Doll & M. Prenzel (Hrsg.), *Bildungsqualität von Schule. Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung* (S. 31–53). Münster: Waxmann.

Krengel, U. (2005). *Einführung in die Wahrscheinlichkeitstheorie und Statistik* (8., erw. Aufl.). Wiesbaden: Vieweg.

Kreutz, M., Hahn, J. & Ackeren, I. v. (2011, Februar). *Zur Wirkung von Post-It-NoteBotschaften bei Fragebogenstudien,* Bamberg.

Kröner, S., Sparfeldt, J. R., Buch, S. R., Zeinz, H. & Rost, D. H. Leistungsangst bei (Lehramts-)Studierenden - Exploration der Zusammenhänge mit den Big Five. *Psychologie in Erziehung und Unterricht,* 187–199.

Kühle, B. (2010). *Zentrale Lernstandserhebungen - ergebnisorientierte Unterrichtsentwicklung? Schulische Strategien beim Umgang mit Ergebnissen aus den*

Schulrückmeldungen im Kontext der ersten Lernstandserhebungen 2004/2005 in Nordrhein-Westfalen (1. Aufl.). Berlin: Köster.

Kühle, B. & Ackeren, I. v. (2012). Wirkungen externer Evaluationsformen für eine evidenzbasierte Schul- und Unterrichtsentwicklung. In U. Bauer, U. Bittlingmayer & A. Scherr (Hrsg.), *Handbuch Bildungs- und Erziehungssoziologie .* Wiesbaden: VS Verlag für Sozialwissenschaften.

Kühle, B. & Peek, R. (2007). Lernstandserhebungen in Nordrhein-Westfalen. Evaluationsbefunde zur Rezeption und zum Umgang mit Ergebnisrückmeldungen in Schulen. *Empirische Pädagogik, 21* (4), 428–447.

Kühn, S. M. (2010). *Steuerung und Innovation durch Abschlussprüfungen?* (Educational Governance, Bd. 11). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kulik, J. A., Bangert-Drowns, R. L. & Kulik, C. L. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 92* (5), 179–188.

Kunter, M. (2005). *Multiple Ziele im Mathematikunterricht* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 51). Münster: Waxmann (Freie Univ., Diss.--Berlin, 2004.).

Kunter, M., Dubberke, T., Baumert, J., Blum, W., Brunner, M. & Jordan, A. (2006). Mathematikunterricht in den PISA-Klassen 2004: Rahmenbedingungen, Formen und Lehr-Lernprozesse. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (S. 161–194). Münster: Waxmann.

Kunter, M. & Klusmann, U. (2010). Kompetenzmessung bei Lehrkräften - Methodische Herausforderungen. *Unterrichtswissenschaft, 38* (1), 68–86.

Kuper, H. (2001). Organisationen im Erziehungssystem. Vorschläge zu einer systemtheoretischen Revision des erziehungswissenschaftlichen Diskurses über Organisation. *Zeitschrift für Erziehungswissenschaft, 4* (1), 83–106.

Kuper, H. & Hartung, V. (2007). Überzeugungen zur Verwendung des Wissens aus Lernstandserhebungen. Eine professionstheoretische Analyse. *Zeitschrift für Erziehungswissenschaft, 10* (2), 214–229.

Kussau, J. (2007). Dimensionen der Koordination: Hierarchische Beobachtung in einer antagonistischen Kooperationsbeziehung. In J. Kussau & T. Brüsemeister (Hrsg.), *Governance, Schule und Politik. Zwischen Antagonismus und Kooperation* (1. Aufl., S. 155–220). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kussau, J. & Brüsemeister, T. (Hrsg.). (2007). *Governance, Schule und Politik. Zwischen Antagonismus und Kooperation* (1. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Lamprecht, M. & Rürup, M. (2012). Bildungsforschung im Rahmen einer evidence based policy: Das Beispiel "Schulinspektion". In A. Wacker, U. Maier & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung: Empirische Befunde und forschungsmethodische Implikationen* (Educational Governance, S. 57–78). Wiesbaden: VS Verl. für Sozialwiss.

Landesamt für Datenverarbeitung und Statistik. (2008). *Statistische Berichte: Allgemeinbildende Schulen in Nordrhein-Westfalen 2007.* Zugriff am 21.03.2011. Verfügbar unter https://webshop.it.nrw.de/gratis/B119%20200700.pdf.

Landesamt für Datenverarbeitung und Statistik. (2009). *Statistische Berichte: Allgemeinbildende Schulen in Nordrhein-Westfalen 2008.* Zugriff am 21.03.2011. Verfügbar unter https://webshop.it.nrw.de/gratis/B119%20200800.pdf.

Landesamt für Datenverarbeitung und Statistik. (2010). *Statistische Berichte: Allgemeinbildende Schulen in Nordrhein-Westfalen 2009.* Zugriff am 21.03.2011. Verfügbar unter https://webshop.it.nrw.de/gratis/B119%20200900.pdf.

Landmann, M. & Schmitz, B. (Hrsg.). (2007). *Selbstregulation erfolgreich fördern. Praxisnahe Trainingsprogramme für effektives Lernen*. Stuttgart: Kohlhammer.

Lehmann-Grube, S. & Nickolaus, R. (2009). Professionalität als kognitive Disposition. In O. Zlatkin-Troitschanskaia, K. Beck, D. Ser, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 59–70). Weinheim: Beltz.

Lehr, D., Hillert, A. & Keller, S. (2009). What can balance the effort? Associations between effort-reward imbalance, overcommitment, and affective disorders in German teachers. *International Journal of Occupational and Environmental Health, 15,* 374–384.

Leuchter, M., Pauli, C., Reusser, K. & Klieme, E. (2006). Unterrichtsbezogene Überzeugungen und handlungsleitende Kognitionen von Lehrpersonen. *Zeitschrift für Erziehungswissenschaft, 9* (4), 562–579.

Leuchter, M., Reusser, K., Pauli, C. & Klieme, E. (2008). Zusammenhänge zwischen unterrichtsbezogenen Kognitionen und Handlungen von Lehrpersonen. In M. Gläser-Zikuda & J. Seifried (Hrsg.), *Lehrerexpertise - Analyse und Bedeutung unterrichtlichen Handelns.* Münster: Waxmann.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl., Studienausg.). Weinheim: Beltz Psychologie-Verl.-Union.

Lind, G. (2009). Amerika als Vorbild? Erwünschte und unerwünschte Folgen aus Evaluationen. In T. Bohl & H. Kiper (Hrsg.), *Lernen aus Evaluationsergebnissen. Verbesserungen planen und implementieren* (S. 61–79). Bad Heilbrunn: Klinkhardt.

Linn, R. L. (2004). Accountability models. In S. Fuhrman & R. F. Elmore (Hrsg.), *Redesigning accountability systems for education* (Critical issues in educational leadership series, S. 73–95). New York, NY: Teachers College Press.

Lipowsky, F. (2003). *Wege von der Hochschule in den Beruf. Eine empirische Studie zum beruflichen Erfolg von Lehramtsabsolventen in der Berufseinstiegsphase*. Bad Heilbrunn/Obb.: Julius Klinkhardt.

Lipowsky, F. (2010). Empirische Befunde zur Wirksamkeit von Lehrerfortbildung. In F. H. Müller, A. Eichenberger, M. Lüders & J. Mayr (Hrsg.), *Lehrerinnen und Lehrer lernen. Konzepte und Befunde zur Lehrerfortbildung* (S. 51–70). Münster: Waxmann.

Maag Merki, K. (2010). Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (Educational Governance, Bd. 7, 1. Aufl., S. 145–169). Wiesbaden: VS Verlag für Sozialwissenschaften.

Maag Merki, K. & Holmeier, M. (2008). Die Implementation zentraler Abiturprüfungen. Erste Ergebnisse zu den Effekten der Einführung auf das schulische Handeln der Lehrpersonen. In E.-M. Lankes (Hrsg.), *Pädagogische Professionalität als Gegenstand empirischer Forschung .* Münster: Waxmann.

Maag Merki, K., Holmeier, M., Jäger, D. & Oerke, B. (2010). Die Effekte der Einführung zentraler Abiturprüfungen auf die Unterrichtsgestaltung in Leistungskursen in der gymnasialen Oberstufe. *Unterrichtswissenschaft, 38* (2), 173–192.

Maier, U. (2008a). Rezeption und Nutzung von Vergleichsarbeiten aus Perspektive von Lehrkräften. *Zeitschrift für Pädagogik, 54* (1), 95–117.

Maier, U. (2008b). Vergleichsarbeiten im Vergleich - Akzeptanz und wahrgenommener Nutzen standardbasierter Leistungsmessung in Baden-Württemberg und Thüringen. *Zeitschrift für Erziehungswissenschaft, 11* (3), 453–474.

Maier, U. (2009). *Wie gehen Lehrerinnen und Lehrer mit Vergleichsarbeiten um? Eine Studie zu testbasierten Schulreformen in Baden-Württemberg und Thüringen* (Schul- und Unterrichtsforschung, Bd. 7). Baltmannsweiler: Schneider Verlag Hohengehren.

Maier, U. (2010a). Accountability policies and teachers' acceptance and usage of school performance feedback - a comparative study. *School effectiveness and school improvement, 21* (2), 145–165.

Maier, U. (2010b). Effekte testbasierter Rechenschaftslegung auf Schule und Unterricht. *Zeitschrift für Pädagogik, 56* (1), 112–128.

Maier, U., Metz, K., Bohl, T., Kleinknecht, M. & Schymala, M. (2012). Vergleichsarbeiten als Instrument der datenbasierten Schul- und Unterrichtsentwicklung in Gymnasien. In A. Wacker, U. Maier & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung. Empirische Befunde und forschungsmethodische Implikationen* (S. 197–224). Wiesbaden: VS Verlag für Sozialwissenschaften.

Maritzen, N. (2008). Schulinspektion. Zur Transformation von Governance-Strukturen im Schulwesen. *Die Deutsche Schule, 100,* 85–96.

Marsh, H. W. (2005). Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology. *The British Psychological Society. Vernon-Wall Lecture* (25).

Maslach, C. & Leiter, M. P. (2006). Teacher burnout: A research agenda. In R. Vandenberghe & A. M. Huberman (Hrsg.), *Understanding and preventing teacher burnout. A sourcebook of international research and practice* (Digitally print. 1. paperback version., S. 295–303). Cambridge: Cambridge Univ. Press.

Maslach, C., Schaufeli, W, & Leiter, M. P. (2001). Job burnout. Annual Review of Psychology, 52, 387-422.

McCrae, R. R. & Costa, P. T. (2006). *Personality in adulthood. A five-factor theory perspective*. New York NY: Guilford Press.

Merzyn, G. (2006). Ideale junger Lehrer und Wirklichkeit. *Plus Lucis* (1//2), 3-6.

Messick, S. J. A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89* (2), 191–216.

Meyer, H. (2009). *Was ist guter Unterricht?* (6. Aufl.). Berlin: Cornelsen-Scriptor.

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. *Zentrale Lernstandserhebungen in der Jahrgangsstufe 8 im Schuljahr 2007/08 – Informationen für die Eltern.* Zugriff am 20.06.2008. Verfügbar unter http://www.standardsicherung.schulministerium.nrw.de/lernstand8/upload/download/mat_07-08/Elterninfo_Lernstandserhebungen-8.pdf.

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2007). Zentrale Lernstandserhebungen (Vergleichsarbeiten) Runderlass des Ministeriums für Schule und Weiterbildung. Verfügbar unter http://www.standardsicherung.schulministerium.nrw.de/lernstand8/upload/download/mat_2006/Erlass_Lernstand_ber_03_07.pdf.

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2010a, 16. Februar). *Information für Lehrerinnen und Lehrer: Vorbereitung der Schülerinnen und Schüler.* Zugriff am 24.07.2011. Verfügbar unter http://www.standardsicherung.schulministerium.nrw.de/lernstand8/lehrer/vorbereitung/.

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2010b, 24. November). *Information für Eltern und Schüler.* Zugriff am 24.07.2011. Verfügbar unter http://www.standardsicherung.schulministerium.nrw.de/lernstand8/eltern/.

Mons, N. (August 2009). *Theoretical and real effects of standardised assessment. Background paper to the study National Testing of Pupils in Europe: Objectives, Organisation and Use of Results EACEA; Eurydice* (Eurydice Network, Hrsg.). Verfügbar unter http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/111EN.pdf.

Moosbrugger, H. (2008a). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 215–259). Heidelberg: Springer.

Moosbrugger, H. (2008b). Klassische Testtheorie. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. S. 110-112). Heidelberg: Springer.

Moschner, B. & Dickhaus, O. (2006). Selbstkonzept. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (Schlüsselbegriffe, 3., überarb. und erw. Aufl., S. 760–767). Weinheim: Beltz PVU.

Müller, A. (2008). *Feedback. Wirkmechanismen und Einsatz in der Personalentwicklung*. Saarbrücken: VDM Verlag Dr. Müller.

Müller, A. (2010). *Rückmeldungen nach Vergleichsarbeiten im Kontext des schulischen Qualitätsmanagements. Drei explorative Studien zu Gestaltung und Rezeption im Anschluss an KOALA-S*. Berlin: Mensch & Buch (Diss.--Ludwig-Maximilians-Universität München, 2009.).

Müller, A. & Hahn, J. (2011, September). *Funktionales und inhaltliches Verständnis von Rückmeldungen nach Vergleichsarbeiten. Eine quasi-experimentelle Studie in Anlehnung an VERA in Nordrhein-Westfalen,* Klagenfurt.

Müller, C., Felbrich, A. & Blömeke, S. (2008). Überzeugungen zum Lehren und Lernen von Mathematik. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und -refendare - erste Ergebnisse zur Wirksamkeit der Lehrerausbildung.* Münster: Waxmann.

Musch, J., Brockhaus, R. & Bröder, A. (2002). Ein Inventar zur Erfassung von zwei Faktoren sozialer Erwünschtheit. *Diagnostica* (48), 121–129.

Musch, J. & Bröder, A. (1999). Ergebnisabhängig asymmetrisches Attributionsverhalten: Motivationale Verzerrung oder rationale Informationsverarbeitung? *Zeitschrift für Sozialpsychologie, 30* (4), 246–254.

Müthing, K. (2005). *Der Einfluss von Selbst- und Fremdbild auf die Verarbeitung von persönlichkeitsbezogenem Feedback*. unveröffentlicht

Nachtigall, C. & Jantowski, A. (2007). Die Thüringer Kompetenztests unter besonderer Berücksichtigung der Evaluationsergebnisse zum Rezeptionsverhalten. *Empirische Pädagogik, 21* (4), 401–410.

Narciss, S. (2006). *Informatives tutorielles Feedback. Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 56). Münster: Waxmann.

Nerdinger, F. W., Blickle, G. & Schaper, N. (2008). *Arbeits- und Organisationspsychologie*. Heidelberg: Springer-Medizin-Verl.

Neuweg, G. H. (2008). Funktion von Aufgaben. In J. Thonhauser (Hrsg.), *Aufgaben als Katalysatoren von Lernprozessen. Eine zentrale Komponente organisierten Lehrens und Lernens aus der Sicht von Lernforschung, allgemeiner Didaktik und Fachdidaktik* (S. 83–98). Münster: Waxmann.

Neuweg, G. H. (2010). Fortbildung im Kontext eines phasenübergreifenden Gesamtkonzepts der Lehrerbildung. In F. H. Müller, A. Eichenberger, M. Lüders & J. Mayr (Hrsg.), *Lehrerinnen und Lehrer lernen. Konzepte und Befunde zur Lehrerfortbildung* (S. 35–49). Münster: Waxmann.

Nichols, S. L. & Berliner, D. C. (2007). *Collateral damage – how high-stake testing corrupts American`s schools*. Cambridge, Massachusetts: Harvard Educational Press.

Nichols, S. L. & Berliner, D. C. (2007a). The pressure of cheat in a high-stake testing environment. In E. M. Anderman & T. B. Murdock (Hrsg.), *Psychology of academic cheating* (S. 289–311). Burlington, Mass.: Elsevier Academic Press.

O`Day, J. A. (2004). Complexity, accountability and school improvement. In S. Fuhrman & R. F. Elmore (Hrsg.), *Redesigning accountability systems for education* (Critical issues in educational leadership series, S. 15–43). New York, NY: Teachers College Press.

Oerke, B., Maag Merki, K., Holmeier, M. & Jäger, J. D. (2011). Changes in student attributions due to the implementation of central exit exams. *Educational Assessment, Evaluation and Accountability, 23,* 223–241.

Opfer, V. D., Pedder, D. J. & Lavicza, Z. (2011). The influence of school orientation to learning on teachers' professional learning chance. *School effectiveness and school improvement, 22* (2), 193–214.

Orth, G. (2005). Bilanz und Ausblick – Lernstandserhebungen in Klasse 9 in NRW. *Schulmagazin 5 bis 10, 10,* 5–8.

Oser, F. & Baeriswyl, F. (2001). Choreographies of Teaching: Bridging Instruction to Learning. In V. Richardson (Hrsg.), *Handbook of Research on Teaching* (4. Auflage, S. 1031–1065). Washington: American Educational Research Association.

Oser, F., Heinzer, S. & Salzmann, P. (2010). Die Messung der Qualität von professionellen Kompetenzprofilen von Lehrpersonen mit Hilfe der Einschätzung von Filmvignetten. *Unterrichtswissenschaft, 38* (1), 4–28.

Otterman, S. (2011, 24. September). State says it analyzed test erasures for cheating: 62 schools proved suspect. *New York Times,* S. A16. Zugriff am 23.10.2011. Verfügbar unter http://www.nytimes.com/2011/09/24/nyregion/in-reversal-new-york-state-says-it-used-erasure-analysis-to-detect-cheating.html?_r=1&scp=9&sq=testing%20high%20school&st=cse.

Parveva, T., Coster, I. de & Noorani, S. (2009). *Nationale Lernstandserhebungen von Schülern in Europa. Ziele, Aufbau und Verwendung der Ergebnisse*. Brüssel: Amt für Veröff.; Eurydice.

Peek, R. (2006). Dateninduzierte Schulentwicklung. In H. Buchen & H.-G. Rolff (Hrsg.), *Professionswissen Schulleitung .* Weinheim: Beltz.

Peek, R. (2007). Wie aussagekräftig sind zentrale Tests? Über den Umgang mit Individualergebnissen aus Schulleistungsstudien, Lernstandserhebungen und zentralen Prüfungen. *forum schule* (1), S. 8-11.

Peek, R., Pallack, A., Dobbelstein, P., Fleischer, J. & Leutner, D. (2006). Lernstandserhebungen 2004 in Nordrhein-Westfalen – zentrale Testergebnisse und Perspektiven für die Schul- und Unterrichtsforschung. In F. Eder, A. Gastager & F. Hofmann (Hrsg.), *Qualität durch Standards? Beiträge zum Schwerpunktthema der 67. Tagung der AEPF.* (S. 187–199). Münster: Waxmann.

Popham, W. J. (2001). Teaching to the test. *Educational leadership, 58* (6), 16–20.

Porst, R. (2009). *Fragebogen. Ein Arbeitsbuch* (Studienskripten zur Soziologie, 2. Auflage.). Wiesbaden: VS Verlag für Sozialwissenschaften / GWV Fachverlage GmbH Wiesbaden. Verfügbar unter http://dx.doi.org/10.1007/978-3-531-91840-2.

Powers, D. E. (1985). Effects of coaching on GRE aptitude test scores. *Journal of Educational Measurement, 22* (2), 121–136.

Powers, D. E. (1988). *Preparing for the SAT: A survey of programs and resources*. New York: College Board Report.

Powers, D. E. (1998). Preparing for the SAT I Reasoning Test - an update. *College Board Report, 98* (5).

Powers, D. E. R. D. A. (1999). Effects of coaching on SAT I: reasoning test scores. *Journal of Educational Measurement, 36* (2), 93–118.

Preiser, S. (2006). Kontrollüberzeugung. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (Schlüsselbegriffe, 3., überarb. und erw. Aufl.). Weinheim: Beltz PVU.

Prieto, L. L., Soria, M. S., Martínez, L. M. & Schaufeli, W. B. (2008). Extension of the job demands-resources model in the prediction of burnout and engagement among teachers over time. *Psicothema, 20* (3), 354–360.

Rakoczy, K. (2007, August). *Videoanalysen in der Unterrichtsforschung,* Ludwigsfelde. Zugriff am 26.03.2011. Verfügbar unter http://www.dipf.de/de/pdf-dokumente/dipf-services/services-bildungsforschung/videoanalysen-in-der-unterrichtsforschung.

Rambow, R. & Bromme, R. (2000). Was SCHÖNs "reflective practitioner" durch Kommunikation mit Laien lernen können. In G. H. Neuweg (Hrsg.), *Wissen - Können - Reflexion. Ausgewählte Verhältnisbestimmungen* (S. 245–263). Innsbruck: Studien-Verl.

Rauin, U. (2007). Im Studium wenig engagiert - im Beruf schnell überfordert. Studienverhalten und Karriere im Lehrerberuf - Kann man Risiken schon im Studium prognostizieren? *Forschung Frankfurt* (3), 60–64.

Raymond, M. E. & Hanushek, E. A. (2003). High-Stake research. *Education Next, 3* (3), 48–55.

Rheinberg, F. & Salisch, M. von. (2008). *Motivation* (Kohlhammer-Urban-Taschenbücher, Bd. 555, 7., aktualisierte Aufl.). Stuttgart: Kohlhammer.

Rindermann, H. (2006). Was messen internationale Schulleistungsstudien? Schulleistungen, Schülerfähigkeiten, kognitive Fähigkeiten, Wissen oder allgemeine Intelligenz? *Psychologische Rundschau, 57* (2), 69–86.

Rolff, H. G. (1993). *Wandel durch Selbstorganisation. Theoretische Grundlagen und praktische Hinweise für eine bessere Schule*. Weinheim und Basel: Juventa Verlag.

Rolff, H. G. (2006). Schulentwicklung, Schulprogramm und Steuergruppe. In H. Buchen & H.-G. Rolff (Hrsg.), *Professionswissen Schulleitung* (S. 296–364). Weinheim: Beltz.

Rolff, H. G. (2007). Zwei Linien der Steuerung der Qualität von Schulen? In H. G. Rolff (Hrsg.), *Studien zu einer Theorie der Schulentwicklung* (Beltz-Bibliothek, S. 195–221). Weinheim: Beltz.

Rosenshine, B. (2003). High-Stakes Testing: Another Analysis. *Education Policy Analysis Archives, 11* (24), 1–8. Zugriff am 06.10.2011.

Ross, S. A. (1973). The economic theory of agency: The principal's problem. *American Economic Association, 63* (2), 134–139.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber Verlag.

Rudolph, M. (2009). Nachhilfe. In K. H. Arnold, U. Sandfuchs & J. Wiechmann (Hrsg.), *Handbuch Unterricht* (2. aktualisierte Aufl., S. 276–279). Stuttgart: UTB.

Ryan, A. M. & Brown, K. W. (2007). Legislating Competence. High-stakes testing policies and their relations with psychological theories and research. In A. J. Elliot & C. S. Dweck (Hrsg.), *Handbook of competence and motivation* (Paperback ed., S. 354–372). New York, NY: Guilford Press.

Ryan, K. E., Ryan, A. M., Arbuthnot, K. & Samuels, M. (2007). Students´ motivation for standardized math exams. *Educational Researcher, 36* (5), 5–13.

Ryan, R. M. & Sapp, A. (2005). Zum Einfluss testbasierter Reformen: High Stake Testing (HST). Motivation und Leistung aus Sicht der Selbstbestimmungstheorie. *Unterrichtswissenschaft, 33* (2), 143–159.

Sacher, W. (2009). *Leistungen entwickeln, überprüfen und beurteilen. Bewährte und neue Wege für die Primar- und Sekundarstufe* (5., überarb. und erw. Aufl.). Bad Heilbrunn: Klinkhardt.

Saum-Aldehoff, T. (2007). *Big Five. Sich selbst und andere erkennen*. Düsseldorf: Patmos.

Schaarschmidt, U. (2002). Die Belastungssituation von Lehrerinnen und Lehrern. Pädagogik, 54 (7/8), 8-13.

Schaarschmidt, U. (2005). *Halbtagsjobber? Psychische Gesundheit im Lehrerberuf - Analyse eines veränderungsbedürftigen Zustandes* (2. Aufl.). Weinheim: Beltz.

Schaarschmidt, U. (2009). Beanspruchung und Gesundheit im Lehrberuf. In O. Zlatkin-Troitschanskaia, K. Beck, D. Ser, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 605–616). Weinheim: Beltz.

Schaarschmidt, U. & Fischer, A. W. (1997). AVEM - ein diagnostisches Instrument zur Differenzierung von Typen gesundheitsrelevanten Verhaltens und Erlebens gegenüber der Arbeit. *Zeitschrift für Differentielle und Diagnostische Psychologie, 18* (3), 151–163.

Schaarschmidt, U. & Fischer, A. W. (2008). *Arbeitsbezogene Verhaltens- und Erlebensmuster (AVEM)*. Frankfurt a. M.: Pearson.

Schaarschmidt, U., Kieschke, U. & Fischer, A. W. (1999). Beanspruchungsmuster im Lehrerberuf. *Psychologie in Erziehung und Unterricht* (46), 244–268.

Schaufeli, W. B. & Bakker, A. B. (2004). Job demands, job resources, and their relationship with burnout and engagement: a multi-sample study. *Journal of Organizational Behavior, 25,* 293–315.

Schaufeli, W. & Bakker, A. (2003). *Utrechter Work Engagement Scale (UWES),* Utrecht University. Zugriff am 05.10.2010. Verfügbar unter http://www.schaufeli.com/downloads/tests/Test%20manual%20UWES.pdf.

Schelten, A. (2009). Lehrerpersönlichkeit - ein schwer fassbarer Begriff. *Die berufsbildende Schule, 61* (2), 39–40.

Schermelleh-Engel, K. & Werner, C. (2008). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. S. 113-133). Heidelberg: Springer.

Scherm, M. (2002). 360°-Feedback für Schulleiter. Chancen und Kritik eines Instruments der Kompetenzentwicklung. *Schulmanagement* (5), 14–16.

Schiefele, U. & Krapp, A. (2000). *Interesse und Lernmotivation. Untersuchungen zu Entwicklung, Förderung und Wirkung ; [Andreas Krapp zu seinem 60. Geburtstag am 3. Juli 2000 gewidmet]*. Münster: Waxmann.

Schildkamp, K., Visscher, A. & Luyten, H. (2009). The effects of the use of a school self-evaluation instrument. *Schhol Effectiveness and School Improvement, 20* (1), 69–88.

Schmitz, B. & Schmidt, M. (2007). Einführung in die Selbstregulation. In M. Landmann & B. Schmitz (Hrsg.), *Selbstregulation erfolgreich fördern. Praxisnahe Trainingsprogramme für effektives Lernen* (S. 9–18). Stuttgart: Kohlhammer.

Schmitz, G. S. (2000). *Zur Struktur und Dynamik der Selbstwirksamkeitserwartung von Lehrern. Ein protektiver Faktor gegen Belastung und Burnout?* Dissertation, Freie Universität Berlin. Berlin. Zugriff am 18.10.2010. Verfügbar unter http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000000315.

Schmotz, C., Felbrich, A. & Kaiser, G. (2010). Überzeugungen angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* (S. 279–305). Münster: Waxmann.

Schneewind, J. (2007). Erfahrungen mit Ergebnisrückmeldungen im Projekt BeLesen - Ergebnisse der Interviewstudie. *Empirische Pädagogik, 21* (4), 368–382.

Schneewind, J. (2007a). *Wie Lehrkräfte mit Ergebnisrückmeldungen aus Schulleistungsstudien umgehen. Ergebnisse aus Befragungen von Berliner Grundschullehrerinnen*. Dissertation, Freie Universität Berlin. Berlin. Zugriff am 18.10.2010. Verfügbar unter http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000002819.

Schöne, C., Dickhäuser, O., Spinath, B. & Stiensmeier-Pelster, J. (2002). *SESSKO. Skalen zur Erfassung des schulischen Selbstkonzepts. Manual*. Göttingen: Hogrefe.

Schrader, F. W. & Helmke, A. (2002). Alltägliche Leistungsbeurteilung durch Lehrer. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (Beltz-Pädagogik, 2., unveränderte, S. 45–58). Weinheim: Beltz.

Schulte, K. (2008). *Selbstwirksamkeitserwartungen in der Lehrerbildung. Zur Struktur und dem Zusammenhang von Lehrer-Selbstwirksamkeitserwartungen, Pädagogischem Professionswissen und Persönlichkeitseigenschaften bei Lehramtsstudierenden und Lehrkräften. Dissertation,* Georg-August-Universität zu Göttingen. Göttingen

Schumacher, L., Paulus, P. & Sieland, B. (2009). Situative Einflussfaktoren auf die Gesundheit und Professionalität von Lehrkräften. In O. Zlatkin-Troitschanskaia, K. Beck, D. Ser, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 617–628). Weinheim: Beltz.

Schwarzer, R. & Jerusalem, M. (2002). Das Konzept der Selbstwirksamkeit. In M. Jerusalem & D. Hopf (Hrsg.), *Selbstwirksamkeit und Motivationsprozesse in Bildungsinstitutionen.* (Zeitschrift für Pädagogik, Beiheft 44, S. 28–53). Weinheim: Beltz.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2004). Standards für die Lehrerbildung: Bildungswissenschaften. Verfügbar unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005). *Bildungsstandards der Kultusministerkonferenz. Erläuterungen zur Konzeption und Entwicklung* (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Hrsg.). München: Luchterhand.

Sembill, D. & Seifried, J. (2009). Konzeptionen, Funktionen und intentionale Veränderungen von Sichtweisen. In O. Zlatkin-Troitschanskaia, K. Beck, D. Ser, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 345–354). Weinheim: Beltz.

Sevincer, A. T. & Oettingen, G. (2009). Ziele. In V. Brandstätter & J. H. Otto (Hrsg.), *Handbuch der Allgemeinen Psychologie - Motivation und Emotion* (S. 37–45). Göttingen: Hogrefe.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *American Educational Research Journal, 15* (4), 4–14.

Shulman, L. S. (1987). Knowledge and Teaching: Foundations of New Reform. *Harvard Educational Review, 57* (1), 1–22.

Siegrist, J., Starke, D., Chandola, T., Godin, I., Marmot, M., Niedhammer, I. et al. (2004). The measurement of effort-reward imbalance at work. European comparisons. *Social Science & Medicine* (58), 1483–1499.

Sjuts, J. (2007, März). *Teaching to the test: Gefahr oder Chance?,* Dortmund. Zugriff am 20.07.2011. Verfügbar unter http://www.mathematik.tu-dortmund.de/ieem/cms/media/BzMU/BzMU2007/Sjuts.pdf.

Sonnentag, S. & Frese, M. (2002). Performance concepts and performance theory. In S. Sonnentag (Hrsg.), *Psychological management of individual performance* (Wiley handbooks in the psychology of management in organizations, S. 3–25). Chichester: Wiley.

Spörl, M. (2009). Motivation durch Zielsetzung. In M. Sauerland (Hrsg.), *Zündstoff Motivation: Motivierungsmethoden für Mitarbeiter, Führungskräfte und Organisationen* (Schriftenreihe Schriften zur Arbeits-, Betriebs- und Organisationspsychologie, Bd. 50, S. 11–34). Hamburg: Kovac.

Stamm, M. (2003). *Evaluation und ihre Folgen für die Bildung. Eine unterschätzte pädagogische Herausforderung* (Internationale Hochschulschriften, Bd. 419). Münster: Waxmann (Univ., Habil.-Schr.--Fribourg, 2002.).

Stecher, B. M. (2002). Consequences of large-scale, high stakes testing on school and classroom practice. In L. S. Hamilton, B. M. Stecher & S. P. Klein (Hrsg.), *Making sense of test-based accountability in education* (S. 79–100). Santa Monica, CA: Rand.

Stecher, B. M. & Barron, S. (2001). Unintended consequences of test-based accountability when testing in "milepost" grade. *Educational Assessment, 7* (4), 259–281.

Stecher, B. M., Chun, T. & Barron, S. (2004). The effect of assessment-driven reform on the teaching of writing in Washington State. In L. Cheng, Y. J. Watanabe & A. Curtis (Hrsg.), *Washback in language testing. Research contexts and methods* (S. 53–69). Mahwah, NJ: Erlbaum.

Steins, G. (2009). Widerstand von Lehrern gegen Evaluationen aus psychologischer Sicht. In T. Bohl & H. Kiper (Hrsg.), *Lernen aus Evaluationsergebnissen. Verbesserungen planen und implementieren* (S. 185–195). Bad Heilbrunn: Klinkhardt.

Steyer, R. & Eid, M. (2001). *Messen und Testen* (Springer-Lehrbuch, 2., korrigierte Aufl.). Berlin: Springer.

Stockmann, R. (2006). Qualitätsmanagement und Evaluation im Vergleich. In W. Böttcher, H. G. Holtappels & M. Brohm (Hrsg.), *Evaluation im Bildungswesen. Eine Einführung in*

Grundlagen und Praxisbeispiele (Grundlagentexte Pädagogik, S. 23–38). Weinheim: Juventa Verlag.

Taddicken, M. (2008). *Methodeneffekte bei Web-Befragungen - Einschränkungen der Datengüte durch ein reduziertes Kommunikationsmedium?* Köln: Herbert von Halem Verlag.

Tenorth, H. E. (2006). Professionalität im Lehrerberuf. Ratlosigkeit der Theorie, gelingende Praxis. *Zeitschrift für Erziehungswissenschaft, 9* (4), 521–544.

Terhart, E. (1986). Organisation und Erziehung. Neue Zugangsweisen zu einem alten Dilemma. *Zeitschrift für Pädagogik, 32* (2), 205–223.

Tresch, S. (2007). *Potenzial Leistungstest. Wie Lehrerinnen und Lehrer Ergebnisrückmeldungen zur Sicherung und Steigerung ihrer Unterrichtsqualität nutzen*. Bern: h.e.p. Verlag.

Verhaeghe, G., Vanhoof, J., Valcke, M. & Petegem, P. van. (2010). Using school performance feedback: perceptions of primary school principals. *School effectiveness and school improvement, 21* (2), 167–188.

Verloop, N., Driel, J. v. & Meijer, P. (2001). Teacher knowledge and the knowledge base of teaching. *International Journal of Educational Research, 35,* 441–461.

Visscher, A. J. (2008). The utilization of school performance feedback systems for school improvement. In A. Breiter, A. Lange & E. Stauke (Hrsg.), *School information systems and data-based decision-making. Schulinformationssysteme und datengestützte Entscheidungsprozesse* (S. 23–37). Frankfurt am Main: Lang.

Visscher, A. J. & Coe, R. (2003). School performance feedback systems: Conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement, 14* (3), 321–349.

Volker, M. A. (2006). Reporting effect size estimates in school psychology research. *Psychology in the Schools, 43* (6), 653–672.

Vollmeyer, R. & Rheinberg, F. (2005). A surprising effect of feedback on learning. *Learning and instruction* (15), 589–602.

Wainer, H. (2010). Schrödinger's cat and the conception of probability in item response theory. *CHANCE, 23* (1), 53–56.

Warner, L. M. & Schwarzer, R. (2009). Empirische Befunde zur Beeinflussbarkeit der Lehrer-Selbstwirksamkeit. In O. Zlatkin-Troitschanskaia, K. Beck, D. Ser, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 629–640). Weinheim: Beltz.

Wegener, H., Fromme, J. & Clausen, M. (2011, September). *Kognitive Aktivierung im Unterricht mit hochbegabten und leistungsstarken Schülerinnen und Schülern. Erste Ergebnisse einer videobasierten Evaluationsstudie,* Klagenfurt.

Wegge, J. & Schmidt, K. H. (2009). Die Zielsetzungstheorie im Überblick. In V. Brandstätter & J. H. Otto (Hrsg.), *Handbuch der Allgemeinen Psychologie - Motivation und Emotion* (S. 174–181). Göttingen: Hogrefe.

Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly, 21* (1), 1–19.

Weinert, F. E. (2001). Concept of competence. A conceptual clarification. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and selecting key competencies.* Seattle: Hogrefe & Huber.

Wößmann, L. (2007). International Evidence on School Competition, Autonomy and Accountability: A Review. *Peabody Journal of Education, 82* (2-3), 473–497.

Wößmann, L. (2008). Zentrale Abschlussprüfungen und Schülerleistungen: Individualanalysen anhand von vier internationalen Tests. *Zeitschrift für Pädagogik, 54* (6), 810–826.

Anhang

Wie man einen Schweinebraten zubereitet (Rezept meines Vaters)

Braten gut unter kaltem Wasser abspülen, dann mit einem sauberen Handtuch abtupfen. Eine Gewürzmischung aus Salz, Pfeffer und Paprika herstellen und damit den Braten satt einreiben. In einen Bräter mehrere Esslöffel Öl (gerne Olivenöl) hineingeben und leicht erhitzen, dann darin den Braten ohne Deckel gut anbraten, bis er von allen Seiten schön braun ist, ruhig etwas dunkler, so dass eine schöne Kruste entsteht.

Zwischenzeitlich eine Fleischbrühe aus Würfeln in einem Stieltopf anrühren, 1/4 l reicht, besser ist aber 1/2 l für mehr Sauce. Topf mit dem angebratenen Braten von der Platte nehmen (zischt dann nicht beim Hinzugießen der Brühe) und nach 1-2 Minuten die heiße Fleischbrühe hinzugeben.

Bratentopf wieder auf die Platte stellen und bei kleiner Flamme mit Deckel (Glasdeckel ist besser, da man schön das leichte Köcheln beobachten kann) ca. 1 Std. bis 1 1/2 Std. köcheln lassen, je nach Größe des Bratens. Der Braten ist weich, wenn man mit der Gabel weich hineinstechen kann.

Kurz vor Ende der Bratzeit, die Brühe andicken: mit einem Becher Crème fraîche auf 1/4 l oder auch mit Schmand köcheln lassen ohne Deckel, so dass die Sauce eindicken kann. Um dickere Sauce zu erhalten, 1 Teelöffel Speisestärke in einer Tasse mit kaltem Wasser anrühren und zur Brühe hinzufügen, kurz aufkochen lassen. Weiter köcheln bis die Sauce genügend eingedickt ist.

Man kann auch den Braten vor dem Einköcheln herausnehmen und ihn im vorgeheizten Ofen warm stellen.

Anschreiben

Anschreiben von Prof. Dr. Isabell van Ackeren und Prof. Dr. Wilfried Bos an die Schulleitungen

Anschreiben von Prof. Dr. Isabell van Ackeren und Prof. Dr. Wilfried Bos an die Lehrkräfte

Anschreiben von Jörn Hahn an die Lehrkräfte

Anschreiben von Prof. Dr. Isabell van Ackeren und Prof. Dr. Wilfried Bos an die Schulleitungen (Nachfassung)

Anschreiben von Prof. Dr. Isabell van Ackeren und Prof. Dr. Wilfried Bos an die Lehrkräfte (Nachfassung)

ag bifo

IFS Institut für Schulentwicklungs-forschung

Prof. Dr. Isabell van Ackeren
Prof. Dr. Wilfried Bos

Kommunikation
Fon 0201-183…
bzw.
Fon 0231 755…

Essen und Dortmund, 05.04.10

Studie über die Vorbereitung auf die Lernstandserhebungen 8/Vergleichsarbeiten 8

Sehr geehrte Frau ,

in den letzten Jahren hat es gravierende Veränderungen im Schulbereich gegeben. In diesem Kontext wurden auch die Lernstandserhebungen 8/Vergleichsarbeiten 8 eingeführt. Verschiedene Studien zum Umgang mit Lernstandserhebungen, die zur Kontrolle des Lernerfolgs und als Unterstützungsinstrument für Lehrerinnen und Lehrer konzipiert wurden, zeigen eine große Unzufriedenheit vieler Lehrerinnen und Lehrer mit dem Verfahren und dem damit verbundenen Aufwand.

Die Untersuchung unseres Promovierenden Jörn Hahn nimmt mit der Vorbereitung auf Lernstandserhebungen einen entscheidenden Abschnitt des Verfahrens in den Blick, der die Funktionalität von Lernstandserhebungen als Unterstützungsinstrument bestimmt. Wir halten die Studie von Herrn Hahn für äußerst wichtig und erkenntniseinträglich und finanzieren das Projekt daher völlig selbstständig aus Haushaltsmitteln der AG Bildungsforschung (Universität Duisburg-Essen) und des Instituts für Schulentwicklungsforschung (Technischen Universität Dortmund).

Wir möchten auch Sie bitten, diese Untersuchung zu unterstützen und die Fragebögen an Ihre Kolleginnen und Kollegen zu verteilen.

Mit freundlichen Grüßen

(Isabell van Ackeren) (Wilfried Bos)

IFS Institut für Schulentwicklungsforschung

Prof. Dr. Isabell van Ackeren
Prof. Dr. Wilfried Bos

Kommunikation
Fon 0201-183…
bzw.
Fon 0231 755…

Essen und Dortmund, 05.04.10

Studie über die Vorbereitung auf die Lernstandserhebungen 8/Vergleichsarbeiten 8

Sehr geehrte Lehrerinnen und Lehrer,

in den letzten Jahren hat es gravierende Veränderungen im Schulbereich gegeben. In diesem Kontext wurden auch die Lernstandserhebungen 8/Vergleichsarbeiten 8 eingeführt. Verschiedene Studien zum Umgang mit Lernstandserhebungen, die zur Kontrolle des Lernerfolgs und als Unterstützungsinstrument für Lehrerinnen und Lehrer konzipiert wurden, zeigen eine große Unzufriedenheit vieler Lehrerinnen und Lehrer mit dem Verfahren und dem damit verbundenen Aufwand.

Die Untersuchung unseres Promovierenden Jörn Hahn nimmt mit der Vorbereitung auf Lernstandserhebungen einen entscheidenden Abschnitt des Verfahrens in den Blick, der die Funktionalität von Lernstandserhebungen als Unterstützungsinstrument bestimmt. Wir halten die Studie von Herrn Hahn für äußerst wichtig und erkenntniseinträglich und finanzieren das Projekt daher völlig selbstständig aus Haushaltsmitteln der AG Bildungsforschung (Universität Duisburg-Essen) und des Instituts für Schulentwicklungsforschung (Technischen Universität Dortmund).

Wir möchten auch Sie bitten, diese Untersuchung zu unterstützen und die Fragebögen an Ihre Kolleginnen und Kollegen zu verteilen.

Mit freundlichen Grüßen

(Isabell van Ackeren) (Wilfried Bos)

Universität Duisburg-Essen
- Campus Essen / Weststadttürme -
Fakultät für Bildungswissenschaften
- Jörn Hahn -
45117 Essen

hahn@ifs.tu-dortmund.de

Institut für Schulentwicklungsforschung

Studie über die Vorbereitung auf die Lernstandserhebungen 8/Vergleichsarbeiten 8

Essen und Dortmund, 05.04.10

Sehr geehrte Lehrerinnen und Lehrer,

im Rahmen meines Forschungs- und Promotionsprojektes möchte ich die Wirkung von LSE auf Unterricht und Schule untersuchen. Erwartet werden Erkenntnisse, mit denen mögliche **unbeabsichtigte Begleiterscheinungen** dieses Instruments aufgezeigt und **uneffektiver Mehraufwand vermieden** werden können.

Worum es geht:
Die Studie beinhaltet die **Befragung aller Mathematiklehrkräfte**, die im aktuellen Halbjahr an einem Gymnasium in Nordrhein-Westfalen **in der achten Jahrgangsstufe** Mathematik unterrichten. Der eingesetzte Fragebogen beinhaltet Fragen zur Vorbereitung auf die LSE sowie allgemeine Einstellungs- und Persönlichkeitsfragen. Ziel ist es, den Umfang und die Art der Vorbereitung zu erfassen. Auch soll geprüft werden, ob Einstellungen und Erlebnisse im Kontext von LSE oder allgemeiner Persönlichkeitsmerkmale die bessere Erklärung für die dabei erlangten Befunde liefern. Der Fragebogen wurde im Vorfeld dem kritischen Urteil von mehr als 30 Lehrkräften unterzogen.

Darum möchte ich Sie bitte:
Es würde mich freuen, wenn Sie mein Projekt unterstützen, indem Sie diesen Fragebogen ausfüllen und **bis zum 02.05.** an mich **zurücksenden**. Das Ausfüllen wird ca. 25 min. in Anspruch nehmen. Für die Rücksendung des Fragebogens benutzen Sie bitte den beigelegten Rückumschlag. Alternativ können Sie den Fragebogen auch online ausfüllen:

www.lernstand.ifs-dortmund.de

(Für jeden online ausgefüllten Fragebogen spende ich 50 Cent an S.O.S. Kinderdorf.)

Die Fragebogenerhebung ist selbstverständlich anonym und alle Daten werden mit größter Sorgfalt verwendet. Jeder Fragebogen ist zur elektronischen Erfassung mit einer **Fragenbogen-Identifikationsnummer** versehen. Diese sind den Fragebögen per Zufall zugeordnet worden und lassen keine Rückschlüsse auf die Schule oder die befragte Person selbst zu.

Ein kleines Dankeschön:
Als kleine symbolische Anerkennung Ihrer Mühe verlose ich unter allen Teilnehmenden 20 Gutscheine für Amazon.de und 20 vom Cornelsen Verlag zur Verfügung gestellte „Mathemagische Momente" je im Wert von 20 €. Für die Teilnahme an der Verlosung ist allerdings die Angabe einer E-Mail-Adresse notwendig, unter der ich Sie ggf. benachrichtigen kann. Wenn Sie am Gewinnspiel teilnehmen wollen und den Fragebogen nicht online ausfüllen möchten, tragen Sie bitte auf dem DinA6-Blatt eine E-Mail-Adresse ein. Wir können Verbesserungsvorschläge nur auf Grundlage vieler ausgefüllter Fragebögen entwickeln. Daher möchte ich Sie herzlich bitten, sich an der Befragung zu beteiligen.

Bei Rückfragen stehe ich jeder Zeit unter obiger E-Mail-Adresse zur Verfügung.

Vielen Dank und freundliche Grüße,

(Jörn Hahn)

IFS Institut für Schulentwicklungsforschung

Prof. Dr. Isabell van Ackeren
Prof. Dr. Wilfried Bos

Kommunikation
Fon 0201-183…
bzw.
Fon 0231 755…

Essen und Dortmund, 05.05.10

Studie über die Vorbereitung auf die Lernstandserhebungen 8 / Vergleichsarbeiten 8

Sehr geehrte Damen und Herren,

vor einigen Wochen erreichte Sie unsere Bitte, die Befragung über die Vorbereitung auf die Lernstandserhebungen zu unterstützen. Wir sind uns sicher, dass Sie unser Anliegen gewissenhaft geprüft und die Fragebögen weitergereicht haben, wenn Sie dies für richtig hielten. Dass viele Schulen bereit sind, sich an diesem Projekt zu beteiligen, zeigt der bisherige Rücklauf nach Ablauf des Stichtages. **Für Ihre Unterstützung bedanken wir uns ganz herzlich!**

Leider ist der Rücklauf bisher für eine repräsentative Datengrundlage gleichwohl noch nicht groß genug. Repräsentative Daten sind aber die Voraussetzung, um die Situation in den Schulen realistisch zu erfassen, angemessene Schlüsse ziehen und sinnvolle Verbesserungsvorschläge erarbeiten zu können. Da vergleichbare Befragungen einen größeren Rücklauf aufweisen, vermuten wir, dass noch mehr Lehrerinnen und Lehrer bereit wären, sich an der Befragung zu beteiligen, das Rücksendedatum für den Fragebogen aber evtl. in Vergessenheit geriet oder zu kurzfristig gewählt war. Wir haben die **Zeit für die Rücksendung** des Fragebogens daher noch einmal **bis zum 17.05.** verlängert.

Wir möchten Sie daher auch noch einmal bitten, das Projekt von Herrn Hahn zu unterstützen und den Mathematiklehrkräften, an die Sie einen Fragebogen weitergereicht haben, auch eines der beiliegenden Schreiben auszuhändigen.

Mit freundlichen Grüßen

(Isabell van Ackeren) (Wilfried Bos)

Essen und Dortmund, 05.05.10

Studie über die Vorbereitung auf die Lernstandserhebungen 8/Vergleichsarbeiten 8

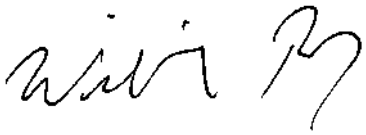
Sehr geehrte Lehrerinnen und Lehrer,

vor einigen Wochen erreichte Sie unser Anschreiben mit der Bitte, die Befragung über die Vorbereitung auf die Lernstandserhebungen zu unterstützen. Sicher haben Sie unser Anliegen gewissenhaft geprüft. Dass viele Lehrerinnen und Lehrer bereit sind, sich an diesem Projekt zu beteiligen, zeigt der bisherige Rücklauf nach Ablauf des Stichtages. **Für Ihre Unterstützung bedanken wir uns ganz herzlich!**

Leider ist der Rücklauf bisher trotzdem für eine repräsentative Datengrundlage noch nicht groß genug. Repräsentative Daten sind aber die Voraussetzung, um die Situation in den Schulen realistisch zu erfassen, angemessene Schlüsse ziehen und sinnvolle Verbesserungsvorschläge erarbeiten zu können. Da vergleichbare Befragungen einen größeren Rücklauf aufweisen, vermuten wir, dass noch mehr Lehrerinnen und Lehrer bereit wären, sich an der Befragung zu beteiligen, das Rücksendedatum für den Fragebogen aber evtl. zu kurzfristig gewählt war oder in Vergessenheit geriet. Wir haben die **Zeit für die Rücksendung** des Fragebogens daher noch einmal **bis zum 17.05.** verlängert.

Wir möchten auch Sie daher bitten, den Fragebogen doch noch auszufüllen bzw. auch nach Ablauf des Rücksendedatums ihn uns sobald wie möglich noch zuzuschicken.

Mit freundlichen Grüßen



(Isabell van Ackeren) (Wilfried Bos)

Fragebögen

Fragebogen Studie A

Fragebogen Studie B

FRAGEBOGEN ÜBER DIE VORBEREITUNG AUF DIE LERNSTANDSERHEBUNGEN

Der vorliegende Fragebogen besteht aus zwei Teilen. In der Regel werden Ihnen in beiden Teilen Aussagen vorgelegt, die zu bewerten sind. Sollten Sie in diesem Halbjahr mehrere achte Klassen in Mathematik unterrichten, wählen Sie bitte eine dieser Klassen aus und beziehen sich bei Ihren Antworten immer nur auf diese. Auch wenn im Fragebogen die maskuline Form verwendet wird, sind immer auch Lehrerinnen, Kolleginnen oder Schülerinnen gemeint.

Für die maschinelle Erfassung Ihrer Antworten ist es notwendig, dass Ihre Wahl jeweils deutlich zu erkennen ist. Sollten Sie einmal ein Kreuz in ein aus Ihrer Sicht falsches Feld gesetzt haben, schwärzen Sie dieses Feld bitte vollständig und setzen das Kreuz in das richtige Feld. ■ ☒ ☐

Zu Beginn benötigen wir einige statistische Angaben:

Wie viele Lehrkräfte unterrichten in diesem Halbjahr in der achten Jahrgangsstufe Ihrer Schule Mathematik? _____ Lehrkräfte

Welcher Altersgruppe gehören Sie an? <36 Jahre ☐ 36-45 Jahre ☐ 46-55 Jahre ☐ 56-65 Jahre ☐

Wie viele Jahre unterrichten Sie bereits Mathematik? ca. _____Jahre

Sie sind... männlich ☐ weiblich ☐

TEIL 1: Die folgenden Aussagen beziehen sich auf die (unmittelbare) Vorbereitung Ihrer Schüler auf die Lernstandserhebungen (LSE):

1. *Wie viele Unterrichtsstunden haben Sie in dieser Klasse ungefähr für die unmittelbare Vorbereitung auf die LSE aufgewendet?* ca. _____ Unterrichtsstunden

2. *Welche der folgenden Optionen haben Sie in dieser Klasse als **unmittelbare Vorbereitung** auf die LSE genutzt? (Mehrfachnennungen möglich)* - Ich habe...

a)	…mit Testaufgaben früherer LSE üben lassen.	◯
b)	…alte Testaufgaben zur Verfügung gestellt.	◯
c)	…mit der Klasse gemeinsam die offizielle Internetseite des Ministeriums/des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) zu den LSE besucht.	◯
d)	…die Schüler auf die offizielle Internetseite des Ministeriums/des IQB hingewiesen.	◯
e)	…die Testsituation simuliert.	◯
f)	…in vorherigen regulären Klassenarbeiten zu den Aufgabenformaten der LSE ähnliche Aufgaben eingebaut.	◯
g)	...in den vorherigen regulären Klassenarbeiten Aufgaben aus vorherigen LSE eingebaut.	◯
h)	…im Unterricht die Beispielaufgaben von der offiziellen Homepage lösen lassen.	◯
i)	…alle Inhaltsbereiche vor den LSE noch einmal wiederholt.	◯
j)	…den Schülern empfohlen, noch einmal alle Inhaltsbereiche zu wiederholen.	◯
k)	…alle Prozessbereiche vor den LSE noch einmal wiederholt.	◯
l)	…den Schülern empfohlen, noch einmal alle Prozessbereiche zu wiederholen.	◯
n)	…eine Prozesskompetenz besonders üben lassen.	◯

o) …keine dieser Optionen genutzt. ❑
Ich habe...

p) …eine hier nicht genannte Option gewählt, nämlich:_____

3. *Welche der folgenden Themen haben Sie im Unterricht angesprochen.* - Ich habe angesprochen...

		mehrfach	einmal	gar nicht
a)	…wie man die Aufgabenstellungen der LSE-Aufgaben richtig versteht.	☐	☐	☐
b)	…wie man den Aufgabenstellungen der LSE-Aufgaben die wichtigen Informationen entnimmt.	☐	☐	☐
c)	…wie man im Hinblick auf die speziellen Antwortformate der LSE richtig antwortet.	☐	☐	☐

4. *Von verschiedenen (Schul-)Buchverlagen gibt es für die drei Fächer Deutsch, Englisch und Mathematik so genannte „Vorbereitungshefte" für die LSE, die jedes Jahr neu aufgelegt werden und speziell auf die aktuell anstehenden LSE zugeschnitten sind.*

		in (fast) allen Stunden	in einigen Stunden	nein
a)	Haben Sie solche Hefte für die unmittelbare Vorbereitung genutzt?	☐	☐	☐
	Wenn Sie diese Vorbereitungshefte eingesetzt haben...	Alle	einige	nein
b)	…haben Sie solche Vorbereitungshefte genutzt, um einige oder alle Inhaltskompetenzen noch einmal zu wiederholen?	☐	☐	☐
c)	…haben Sie solche Vorbereitungshefte eingesetzt, um einige oder alle Prozesskompetenzen noch mal zu wiederholen?	☐	☐	☐
d)	…haben Sie Schüler darauf aufmerksam gemacht, dass es Vorbereitungshefte zu kaufen oder in der Schule zu leihen/ einzusehen gibt?	☐	☐	☐

5. *Weiter gibt es für das Fach Mathematik so genannte „Lernhilfen", Hefte mit Übungsaufgaben, die sich an den Kernlehrplänen orientieren.*

		in (fast) allen	in einigen	nein
a)	Haben Sie solche Hefte in den Unterrichtsstunden unmittelbar vor den LSE genutzt?	☐	☐	☐
	Wenn Sie diese Lernhilfen eingesetzt haben...	Alle	einige	nein
b)	…haben Sie solche Lernhilfen genutzt, um einige oder alle Inhaltskompetenzen noch einmal zu wiederholen?	☐	☐	☐
c)	…haben Sie solche Lernhilfen genutzt, um einige oder alle Prozesskompetenzen noch einmal zu wiederholen?	☐	☐	☐
d)	…haben Sie Schüler darauf aufmerksam gemacht, dass es Lernhilfen zu kaufen oder in der Schule zu leihen/einzusehen gibt?	☐	☐	☐

6. *Bitte geben Sie an, ob Sie folgende Strategien Ihren Schülern mit auf den Weg gegeben haben.* - Ich habe ihnen geraten...

		ja	nein
a)	…sich nicht zu lange mit einer Frage aufzuhalten.	☐	☐
b)	…sich mit dem Antwortformat vertraut zu machen.	☐	☐
c)	…alle Antworten in Betracht zu ziehen, bevor man sich entscheidet.	☐	☐
d)	…die Instruktionen und Fragen genau zu lesen.	☐	☐
e)	…bei Multiple-Choice-Fragen zu raten, wenn man die Antwort nicht weiß.	☐	☐
f)	…zuerst die Fragen zu beantworten, bei denen man sich sicher ist.	☐	☐
g)	…sich spontane Einfälle zu notieren.	☐	☐
h)	…auf die grammatikalische Einschränkung der möglichen Antworten zu achten.	☐	☐

i)	…im Zweifel die erste Idee zu wählen, weil dies meist das Beste ist.	☐	☐

Haben Sie zu einer anderen Strategie geraten?

k) ja, nämlich:_____

7.	*Was denken Sie, wie wichtig sind die Lernstandserhebungen für…*	gar nicht wichtig					sehr wichtig	weiß nicht
a)	…die Mehrheit Ihrer Schüler?	☐	☐	☐	☐	☐	☐	☐
b)	…die Mehrheit deren Eltern?	☐	☐	☐	☐	☐	☐	☐
c)	…die Schulleitung?	☐	☐	☐	☐	☐	☐	☐
d)	…die Mehrheit der anderen Deutsch-, Englisch- bzw. Mathematiklehrkräfte an Ihrer Schule, die aktuell mindestens eines dieser Fächer in einer achten Klasse unterrichten?	☐	☐	☐	☐	☐	☐	☐
e)	…Sie persönlich?	☐	☐	☐	☐	☐	☐	☐

Die nachfolgenden Fragen beziehen sich auf einen längeren Zeitraum:

8. *Wenn Sie sich noch einmal an die Klasse zurückerinnern, in der Sie zum ersten Mal die LSE durchgeführt haben, welche Aussage trifft dann zu? (Bitte nur eine Antwort in diesem Block ankreuzen.)*

a)	Ich bin zum ersten Mal als Lehrer mit LSE konfrontiert.	❑
b)	Ich bereite meine Schüler genauso intensiv auf die LSE vor wie vorher auch.	❑
c)	Ich bereite meine Schüler in diesem Jahr intensiver auf die LSE vor als frühere Jahrgänge.	❑
d)	Ich bereite meine Schüler in diesem Jahr weniger intensiv auf die LSE vor als frühere Jahrgänge.	❑

Wenn Sie in Frage 8 eine der Antwortmöglichkeiten b)-d) gewählt haben:

9. *Welche Gründe gibt es dafür?*

a)	Heute kenne ich mich mit den LSE besser aus.	☐
b)	In den vergangenen Jahren hat der Druck für mich *zugenommen*, bei den LSE gut abzuschneiden.	☐
c)	In den vergangenen Jahren hat der Druck für Schüler *zugenommen*, bei den LSE gut abzuschneiden.	☐
d)	In den vergangenen Jahren hat der Druck für mich *abgenommen*, bei den LSE gut abzuschneiden.	☐
e)	In den vergangenen Jahren hat der Druck für Schüler *abgenommen*, bei den LSE gut abzuschneiden.	☐

f) Andere Gründe, nämlich... _____

10. *Die nachfolgenden Fragen beziehen sich auf das ganze bisherige Schuljahr, nicht nur auf die unmittelbare Vorbereitung vor den LSE:*

		ja	nein
a)	Haben Sie eine Prozesskompetenz in dieser achten Klasse intensiver behandelt als in den Jahrgängen zuvor?	☐	☐
aa)	Falls ja, geschah dies mit Blick auf die aktuellen LSE?	☐	☐
b)	Haben Sie eine Prozesskompetenz in dieser achten Klasse weniger intensiv behandelt als in den Jahrgängen zuvor?	☐	☐
bb)	Falls ja, geschah dies mit Blick auf die aktuellen LSE?	☐	☐
c)	Gab es in der Fachgruppe Mathematik Absprachen, in diesem Schuljahr bestimmte Kompetenzen in der achten Klasse intensiver zu behandeln als in den Jahrgängen zuvor?	☐	☐

		ja	nein
cc)	Falls ja, geschah dies mit Blick auf die aktuellen LSE?	☐	☐
d)	Gab es in der Fachgruppe Mathematik Absprachen, in diesem Schuljahr bestimmte Kompetenzen in der achten Klasse weniger intensiver zu behandeln als in Jahrgängen davor?	☐	☐
dd)	Falls ja, geschah dies mit Blick auf die aktuellen LSE?	☐	☐
e)	Haben Sie einzelnen Schülern empfohlen, sich speziell auf die LSE vorzubereiten?	☐	☐
f)	Haben Sie Schülern empfohlen mit Blick auf die LSE Nachhilfe zu nehmen?	☐	☐
g)	Nehmen Schüler Ihrer Klasse nach Ihrem Kenntnisstand Nachhilfe speziell mit Blick auf die LSE?	☐	☐
h)	Haben Ihre Schüler nach Ihrem Kenntnisstand für die LSE außerhalb des Unterrichts besonders geübt?	☐	☐
i)	Haben Sie der Klasse empfohlen, sich speziell auf die LSE vorzubereiten?	☐	☐

Nun geht es um die Funktionen der LSE:

11.	*Wie stark stimmen Sie folgenden Aussagen zu bzw. lehnen Sie diese ab? Die LSE sind dazu gedacht…*	lehne völlig ab					stimme völlig zu
a)	…den Schülern eine Rückmeldung über ihre mathematischen Kompetenzen zu geben.	☐	☐	☐	☐	☐	☐
b)	…die Lehrer zu unterstützen, einfacher zu einer Zeugnisnote zu gelangen.	☐	☐	☐	☐	☐	☐
c)	…eine Rückmeldung über unser Schulsystem zu geben.	☐	☐	☐	☐	☐	☐
d)	…den Lehrkräften ein Arbeitszeugnis auszustellen.	☐	☐	☐	☐	☐	☐
e)	...für die Schüler eine Generalprobe für zentrale Prüfungen in der weiteren Schulzeit zu sein.	☐	☐	☐	☐	☐	☐
f)	…Lehrkräften Hinweise für die Klassenarbeits-konstruktion zu geben.	☐	☐	☐	☐	☐	☐
g)	…ein Schulranking zu erstellen.	☐	☐	☐	☐	☐	☐
h)	Lehrkräften Hinweise für die weitere Unterrichts-planung zu geben.	☐	☐	☐	☐	☐	☐
i)	…finanzielle und personelle Mittel an Schulen richtig zuzuweisen.	☐	☐	☐	☐	☐	☐

Als nächstes geht es um Ihre Überlegungen, auf die LSE vorzubereiten:

Wenn Sie Ihre Schüler auf die Lernstandserhebungen vorbereitet haben (sonst weiter mit Frage13):

12.		lehne völlig ab					stimme völlig zu
a)	Mir ist es wichtig, dass das Ansehen unserer Schule nicht unter schlechten Ergebnissen bei den LSE leidet.	☐	☐	☐	☐	☐	☐
b)	Ich möchte verhindern, dass der Schule finanzielle Mittel gestrichen werden.	☐	☐	☐	☐	☐	☐
c)	Ich möchte negative Rückmeldungen an mich verhindern, die aufgrund eines schlechten Abschneidens meiner Schüler geäußert werden könnten.	☐	☐	☐	☐	☐	☐
d)	Ich möchte nicht an einer Schule arbeiten, die bei den LSE schlecht abschneidet.	☐	☐	☐	☐	☐	☐

		lehne völlig ab					stimme völlig zu
e)	Ich möchte den Schülern ermöglichen, mit einem guten Ergebnis bei den LSE eine bessere Note zu bekommen.	☐	☐	☐	☐	☐	☐
f)	Ich möchte meinen Schülern die Chance geben, über die LSE Selbstvertrauen zu erlangen.	☐	☐	☐	☐	☐	☐
g)	Ich möchte den Erwartungen meiner Schüler genügen, sie optimal auf die LSE vorzubereiten.	☐	☐	☐	☐	☐	☐
h)	Ich möchte den Erwartungen der Schulleitung genügen, die Schüler optimal auf die LSE vorzubereiten.	☐	☐	☐	☐	☐	☐
i)	Ich möchte meinen Schülern helfen, ihren vollen Kompetenzumfang in den LSE zu zeigen.	☐	☐	☐	☐	☐	☐
j)	Ich möchte meinen Schülern die Angst vor den LSE nehmen.	☐	☐	☐	☐	☐	☐
k)	Ich möchte verhindern, dass Schüler lediglich deshalb eine falsche Antwort ankreuzen, weil sie das Aufgabenformat nicht richtig verstanden haben.	☐	☐	☐	☐	☐	☐
l)	Ich möchte den Erwartungen der Eltern genügen, die Schüler optimal auf die LSE vorzubereiten.	☐	☐	☐	☐	☐	☐

Wenn Sie Ihre Schüler auf die Lernstandserhebungen NICHT vorbereitet haben:

13.		lehne völlig ab					stimme völlig zu
a)	Ich möchte keine wertvolle Unterrichtszeit für die Vorbereitung auf die LSE verwenden.	☐	☐	☐	☐	☐	☐
b)	Ich möchte mich keiner ‚Testhysterie' unterwerfen.	☐	☐	☐	☐	☐	☐
c)	Ich möchte Instrumente wie die LSE nicht durch eine Vorbereitung legitimieren.	☐	☐	☐	☐	☐	☐
d)	Ich möchte mir keine Gedanken über eine spezielle Vorbereitung auf die LSE machen.	☐	☐	☐	☐	☐	☐
e)	Ich möchte verhindern, durch eine zu starke Vorbereitung auf die LSE ihre Aussagekraft der LSE zu gefährden.	☐	☐	☐	☐	☐	☐
f)	Ich möchte verhindern, dass Schülern durch eine Vorbereitung bei den LSE besser abschneiden als sie in Wirklichkeit sind.	☐	☐	☐	☐	☐	☐

TEIL 2: Im zweiten Teil geht es um den Umgang mit Leistung an Ihrer Schule, Ihre Sicht auf den Mathematikunterricht, Ziele Ihres Mathematikunterrichts und einige Einschätzungen zu Ihrer Person.

Zuerst interessiert uns nun, welchen Stellenwert Leistung an Ihrer Schule besitzt:

14.	*Inwieweit treffen folgende Aussagen zu?*	trifft gar nicht zu					trifft völlig zu
a)	Gute schulische Leistungen werden an unserer Schule öffentlich hervorgehoben.	☐	☐	☐	☐	☐	☐
b)	In unserer Schule werden hohe Anforderungen an die Schüler gestellt.	☐	☐	☐	☐	☐	☐
c)	Misserfolge von Schülern werden an unserer Schule von den Lehrkräften verantwortet.	☐	☐	☐	☐	☐	☐
d)	Schulversagen wird bei uns als Problem der Schule und weniger des Schülers gesehen.	☐	☐	☐	☐	☐	☐
e)	Unsere Schule hat den Ehrgeiz, besser zu sein als andere Schulen.	☐	☐	☐	☐	☐	☐
f)	Ich stelle hohe Anforderungen an meine Schüler.	☐	☐	☐	☐	☐	☐
g)	Für die Erfolge meiner Schüler trage ich eine große Verantwortung.	☐	☐	☐	☐	☐	☐
h)	Misserfolge liegen vor allem in der Verantwortung der Schüler selbst.	☐	☐	☐	☐	☐	☐
i)	Ich bemühe mich darum, dass meine Klasse bessere Leistungen erbringt als Parallelklassen.	☐	☐	☐	☐	☐	☐
j)	Ich bemühe mich darum, dass meine Klasse bessere Leistungen erbringt als Schüler anderer Schulen.	☐	☐	☐	☐	☐	☐

Weiter möchten wir gern wissen, wie Sie die Rolle der Mathematik und wie Sie Ihre Rolle im Mathematikunterricht sehen. Einige der folgenden Fragen bzw. Aussagen mögen dabei sehr persönlich erscheinen. Wir möchten Sie dennoch bitten, diese Fragen nach Möglichkeit zu beantworten:

15.	*Wie stark stimmen Sie folgenden Aussagen zu bzw. lehnen Sie diese ab?*	lehne völlig ab					stimme völlig zu
a)	In der Mathematik ist es nicht nur wichtig, die richtige Lösung zu finden, sondern auch zu verstehen, warum diese Lösung richtig ist.	☐	☐	☐	☐	☐	☐
b)	Lehrpersonen sollten Schülern, die Schwierigkeiten mit einer Sachaufgabe haben, die Möglichkeit geben, weiter nach einer Lösung zu suchen.	☐	☐	☐	☐	☐	☐
c)	Lehrpersonen sollten Schülern die Möglichkeit geben, ihre eigenen Wege zu finden, um eine Aufgabe zu lösen.	☐	☐	☐	☐	☐	☐
d)	Schüler können durchaus auch ohne Hilfe der Lehrperson Lösungswege für mathematische Aufgaben finden.	☐	☐	☐	☐	☐	☐
e)	Lehrpersonen sollten Schüler ermutigen, eigene Lösungen für mathematische Aufgaben zu finden, auch wenn diese nicht effizient sind.	☐	☐	☐	☐	☐	☐
f)	Es hilft den Schülern, wenn sie verschiedene Lösungswege für eine bestimmte Aufgabe diskutieren.	☐	☐	☐	☐	☐	☐
g)	Man muss Schülern exakte Verfahren beibringen, damit sie mathematische Probleme lösen können.	☐	☐	☐	☐	☐	☐

		lehne völlig ab					stimme völlig zu
h)	Um gut in Mathematik zu sein, muss ein Schüler gut zuhören können.	☐	☐	☐	☐	☐	☐
i)	Um gut in Mathematik zu sein, muss man Aufgaben schnell lösen können.	☐	☐	☐	☐	☐	☐
j)	Schüler lernen Mathematik am besten, indem sie den Erklärungen der Lehrperson aufmerksam folgen.	☐	☐	☐	☐	☐	☐
k)	Wenn Schüler besser in Mathematik werden wollen, müssen sie einfach sehr viel üben.	☐	☐	☐	☐	☐	☐

16. *Die nachfolgenden Aussagen beziehen sich auf ihre aktuelle Situation als Lehrkraft. Wie stark würden Sie folgenden Aussagen zustimmen?*

		lehne völlig ab					stimme völlig zu
a)	Wenn ich mir überlege, welche Herausforderungen mir als Mathematiklehrkraft begegnen, halte ich mich für einen guten Pädagogen.	☐	☐	☐	☐	☐	☐
b)	Ich erkenne oft nicht, ob ein Schüler etwas verstanden hat.	☐	☐	☐	☐	☐	☐
c)	Wenn ein Schüler eine Aufgabe falsch löst, ist mir der Grund dafür oft nicht klar.	☐	☐	☐	☐	☐	☐
d)	Die Ursache für einen Schülerfehler sehe ich in der Regel sofort.	☐	☐	☐	☐	☐	☐
e)	Ich bin gut darin, Alltagsbezüge zur Mathematik herzustellen.	☐	☐	☐	☐	☐	☐
f)	Das für eine Klasse richtige Lerntempo treffe ich in der Regel genau.	☐	☐	☐	☐	☐	☐

17. *Die nachfolgenden Aussagen beziehen sich auf das Erklären und Anregen im Mathematikunterricht.*

		lehne völlig ab					stimme völlig zu
a)	Ich traue mir zu, Schüler anzuregen, selbst für komplizierte mathematische Sachverhalte und Probleme Erklärungen und Lösungswege zu finden.	☐	☐	☐	☐	☐	☐
b)	Ich bin mir sicher, mit meinem Unterricht auch bei insgesamt desinteressierten Schülern das Interesse für Mathematik wecken zu können.	☐	☐	☐	☐	☐	☐
c)	Ich traue mir zu, meinen Schülern auch komplizierte Beweise der Schulmathematik verständlich zu erklären.	☐	☐	☐	☐	☐	☐
d)	Ich weiß, dass ich Diskussionen über Lösungswege auch in „trägen Klassen“ anregen könnte, wenn ich das wollte.	☐	☐	☐	☐	☐	☐
e)	Ich weiß, dass es mir gelingen würde, anderen die Verbindungen zwischen zwei mathematischen Bereichen aufzuzeigen, auch wenn der Zusammenhang schwierig zu erkennen ist.	☐	☐	☐	☐	☐	☐
f)	Wenn ich es möchte, gelingt es mir nach wenigen Unterrichtsstunden, dass die Schüler in meinem Unterricht selbstständig Aufgaben bearbeiten und Probleme lösen.	☐	☐	☐	☐	☐	☐
g)	Ich kann einem Schüler auch schwer zu verstehende Bereiche der Mathematik erklären, sodass dieser es versteht.	☐	☐	☐	☐	☐	☐

		lehne völlig ab					stimme völlig zu
h)	Meine Schüler zum Bearbeiten von mathematischen Aufgaben zu bewegen, bleibt für mich ein echtes Problem, wie sehr ich mich auch engagiere.	☐	☐	☐	☐	☐	☐

Nun möchten wir erfassen, wie Sie Ihren Beruf erleben:

18.	*In den folgenden Aussagen geht es um das Engagement für Ihre Arbeit. Bitte geben Sie jeweils an, wie sehr Sie den Aussagen zustimmen.*	lehne völlig ab					stimme völlig zu
a)	Bei meiner Arbeit bin ich voll überschäumender Energie.	☐	☐	☐	☐	☐	☐
b)	Andere wollen immer gerade dann, dass ich etwas für sie tue, wenn ich mich erhole/entspanne.	☐	☐	☐	☐	☐	☐
c)	Beim Arbeiten fühle ich mich fit und tatkräftig.	☐	☐	☐	☐	☐	☐
d)	Ich arbeite nicht gern, wenn ich es tun muss.	☐	☐	☐	☐	☐	☐
e)	Ich bin von meiner Arbeit begeistert.	☐	☐	☐	☐	☐	☐
f)	Wenn man nicht so viel dafür arbeiten müsste, wäre ich gerne erfolgreicher.	☐	☐	☐	☐	☐	☐
g)	Meine Arbeit inspiriert mich.	☐	☐	☐	☐	☐	☐
h)	Wenn ich arbeite, muss ich immer daran denken, was ich jetzt Schöneres tun könnte.	☐	☐	☐	☐	☐	☐
i)	Wenn ich morgens aufstehe, freue ich mich auf meine Arbeit.	☐	☐	☐	☐	☐	☐
j)	Wenn draußen die Sonne scheint, kann ich einfach nicht arbeiten.	☐	☐	☐	☐	☐	☐
k)	Ich fühle mich glücklich, wenn ich intensiv arbeite.	☐	☐	☐	☐	☐	☐
l)	Ich sehe nicht ein, warum ich Dinge auswendig lernen soll, die man nachschauen kann.	☐	☐	☐	☐	☐	☐
m)	Ich bin stolz auf meine Arbeit.	☐	☐	☐	☐	☐	☐
n)	Bei Arbeiten, die ich nicht machen möchte, langweile ich mich sehr schnell.	☐	☐	☐	☐	☐	☐
o)	Ich gehe völlig in meiner Arbeit auf.	☐	☐	☐	☐	☐	☐
p)	Warum soll ich etwas gleich erledigen, wenn dafür morgen auch noch Zeit ist?	☐	☐	☐	☐	☐	☐
q)	Meine Arbeit reißt mich mit.	☐	☐	☐	☐	☐	☐

19.	*Wie sehr stimmen Sie diesen Aussagen bezüglich der Belastung durch Ihre Arbeit zu?*	lehne völlig ab					stimme völlig zu
a)	Neben der beruflichen Tätigkeit bleibt mir noch genug Zeit für Familie und Hobbys.	☐	☐	☐	☐	☐	☐
b)	Wenn ich mir die Freizeit gut einteile, habe ich genug Zeit zum Entspannen.	☐	☐	☐	☐	☐	☐
c)	Meine schulische Belastung schränkt meine Freizeit-aktivitäten fühlbar ein.	☐	☐	☐	☐	☐	☐
d)	Ich glaube, dass mein Beruf meine Gesundheit belastet.	☐	☐	☐	☐	☐	☐
e)	Ich habe das Gefühl, dass ich mit der zeitlichen Belastung des Lehrerberufs nicht fertig werde.	☐	☐	☐	☐	☐	☐
f)	Der berufliche Stress wirkt sich negativ auf mein Privatleben aus.	☐	☐	☐	☐	☐	☐

		lehne völlig ab					stimme völlig zu
g)	Ich fühle mich wegen meiner beruflichen Belastung oft müde und abgespannt.	☐	☐	☐	☐	☐	☐
h)	Bei vielen Dingen, die ich tun muss, bekomme ich ein unangenehmes Gefühl.	☐	☐	☐	☐	☐	☐
i)	Bei Arbeiten, die ich nicht machen möchte, schaffe ich es, sie nicht machen zu müssen.	☐	☐	☐	☐	☐	☐

20. *Die folgenden Aussagen beziehen sich auf Ihre Emotionen bezüglich Ihrer Arbeit. Bitte geben Sie den Grad der Zustimmung an.*

		lehne völlig ab					stimme völlig zu
a)	Beim Arbeiten komme ich leicht in Zeitdruck.	☐	☐	☐	☐	☐	☐
b)	Es passiert mir oft, dass ich schon beim Aufwachen an Arbeitsprobleme denke.	☐	☐	☐	☐	☐	☐
c)	Wenn ich nach Hause komme, fällt mir das Abschalten von der Arbeit sehr leicht.	☐	☐	☐	☐	☐	☐
d)	Ich stelle mich dumm, um unangenehme Arbeiten zu vermeiden.	☐	☐	☐	☐	☐	☐
e)	Diejenigen, die mir am nächsten stehen sagen, ich opfere mich zu sehr für meinen Beruf auf.	☐	☐	☐	☐	☐	☐
f)	Ich warte, bis sich unangenehme Arbeiten von selbst erledigen.	☐	☐	☐	☐	☐	☐
g)	Die Arbeit lässt mich selten los, das geht mir abends noch im Kopf rum.	☐	☐	☐	☐	☐	☐
h)	Wenn ich etwas verschiebe, was ich eigentlich heute tun müsste, kann ich nachts nicht schlafen.	☐	☐	☐	☐	☐	☐

21. *In den folgenden Aussagen geht es um Arbeit als Pflicht. Wie sehr stimmen Sie diesen Aussagen zu?*

		lehne völlig ab					stimme völlig zu
a)	Ich bin immer vorbereitet.	☐	☐	☐	☐	☐	☐
b)	Ich achte auch auf Details.	☐	☐	☐	☐	☐	☐
c)	Meine Hausarbeit mache ich sofort.	☐	☐	☐	☐	☐	☐
d)	Meine Pläne setze ich in die Tat um.	☐	☐	☐	☐	☐	☐
e)	Ich mache Pläne und halte mich daran.	☐	☐	☐	☐	☐	☐
f)	Ich benötige oft mehr Zeit als nötig.	☐	☐	☐	☐	☐	☐
g)	Mir fällt es schwer, mich an die Arbeit zu setzen.	☐	☐	☐	☐	☐	☐
h)	Ich mache gerade so viel wie nötig.	☐	☐	☐	☐	☐	☐
i)	Ich überprüfe meine Arbeit.	☐	☐	☐	☐	☐	☐
j)	Wenn möglich, weiche ich Pflichten aus.	☐	☐	☐	☐	☐	☐

Abschließend haben wir noch einige Fragen zu Ihren Zielen in Ihrem Mathematikunterricht:

22. *Wie sehr spielen diese Ziele für Sie eine Rolle?*
- Ich möchte, dass meine Schüler…

		trifft gar nicht zu					trifft völlig zu
a)	…mathematische Algorithmen und Verfahren kennen.	☐	☐	☐	☐	☐	☐
b)	…in der Lage sind, sich mathematische Fakten, Regeln und Lösungsschritte zu merken.	☐	☐	☐	☐	☐	☐
c)	…rechnerischen Fertigkeiten entwickeln.	☐	☐	☐	☐	☐	☐
d)	…in der Lage sind, mathematische Berechnungen	☐	☐	☐	☐	☐	☐

schnell und fehlerfrei auszuführen.

	Ich möchte, dass meine Schüler…	trifft gar nicht zu					trifft völlig zu
e)	…Hilfsmittel wie Taschenrechner oder Tabellenkalkulationsprogramme sinnvoll anwenden können.	☐	☐	☐	☐	☐	☐
f)	…wissen, wann ein Taschenrechner benutzt werden soll und wann nicht.	☐	☐	☐	☐	☐	☐
g)	…lernen einen mathematischen Essay zu schreiben.	☐	☐	☐	☐	☐	☐
h)	…lernen, mathematische Ideen zu erläutern bzw. in Worte zu fassen.	☐	☐	☐	☐	☐	☐
i)	…in der Lage sind, über mathematische Inhalte zu kommunizieren.	☐	☐	☐	☐	☐	☐
j)	…die mathematische Terminologie korrekt benutzen können.	☐	☐	☐	☐	☐	☐
k)	…lernen, mathematisch zu argumentieren.	☐	☐	☐	☐	☐	☐
l)	…ggf. geeignete Nachschlagewerke, Zeitschriften oder auch Internetsuchmaschinen nutzen.	☐	☐	☐	☐	☐	☐
m)	…die geeignete Möglichkeit kennen und nutzen, um mathematische Lösungen zu präsentieren.	☐	☐	☐	☐	☐	☐
n)	…die logische Struktur der Mathematik verstehen.	☐	☐	☐	☐	☐	☐
o)	…das Gefühl bekommen, dass Mathematik etwas ist, welches sie beherrschen.	☐	☐	☐	☐	☐	☐
p)	…sich für Mathematik zu interessieren beginnen.	☐	☐	☐	☐	☐	☐
q)	…Spaß an Mathematik haben.	☐	☐	☐	☐	☐	☐
r)	…ein Bewusstsein für die Wichtigkeit der Mathematik im täglichen Leben entwickeln.	☐	☐	☐	☐	☐	☐
s)	…das Prinzip des mathematischen Beweises verstehen.	☐	☐	☐	☐	☐	☐
t)	…in der Lage sind, mathematische Aussagen zu beweisen.	☐	☐	☐	☐	☐	☐
u)	…die logische Struktur der Mathematik verstehen.	☐	☐	☐	☐	☐	☐
v)	…lernen, wie verschiedene mathematische Begriffe und Ideen miteinander zusammenhängen.	☐	☐	☐	☐	☐	☐
w)	…lernen, auch neue mathematische Probleme zu lösen.	☐	☐	☐	☐	☐	☐
x)	…selbst Problemlösungen entwickeln.	☐	☐	☐	☐	☐	☐
y)	…Problemlösestrategien vergleichen und bewerten können.	☐	☐	☐	☐	☐	☐
z)	…in der Lage sind, lebensweltliche Problemstellungen mathematisch zu modellieren.	☐	☐	☐	☐	☐	☐
ä)	…mathematische Lösungen am Realmodell überprüfen und ggf. Anpassungen am Modell oder der Lösung vornehmen.	☐	☐	☐	☐	☐	☐
ö)	…zu mathematischen Modellen passende Realsituationen finden können.	☐	☐	☐	☐	☐	☐

Haben Sie noch Anmerkungen zum Fragebogen oder den LSE, dann können Sie die hier notieren:

_____

_____

_____

_____

Vielen herzlichen Dank für Ihre Teilnahme!

FRAGEBOGEN ÜBER DIE VORBEREITUNG AUF DIE LERNSTANDSERHEBUNGEN

Der vorliegende Fragebogen besteht aus zwei Teilen. In der Regel werden Ihnen in beiden Teilen Aussagen vorgelegt, die zu bewerten sind. Sollten Sie in diesem Halbjahr mehrere achte Klassen in Mathematik unterrichten, wählen Sie bitte eine dieser Klassen aus und beziehen sich bei Ihren Antworten immer nur auf diese. Auch wenn im Fragebogen die maskuline Form verwendet wird, sind immer auch Lehrerinnen, Kolleginnen oder Schülerinnen gemeint.

Für die maschinelle Erfassung Ihrer Antworten ist es notwendig, dass Ihre Wahl jeweils deutlich zu erkennen ist. Sollten Sie einmal ein Kreuz in ein aus Ihrer Sicht falsches Feld gesetzt haben, schwärzen Sie dieses Feld bitte vollständig und setzen das Kreuz in das richtige Feld. ■ ☒ ☐

Zu Beginn benötigen wir einige statistische Angaben:

Wie viele Lehrkräfte unterrichten in diesem Halbjahr in der achten Jahrgangsstufe Ihrer Schule Mathematik? _____ Lehrkräfte

Welcher Altersgruppe gehören Sie an? <36 Jahre ☐ 36-45 Jahre ☐ 46-55 Jahre ☐ 56-65 Jahre ☐

Wie viele Jahre unterrichten Sie bereits Mathematik? ca. _____Jahre

Sie sind... männlich ☐ weiblich ☐

TEIL 1: Die folgenden Aussagen beziehen sich auf die (unmittelbare) Vorbereitung Ihrer Schüler auf die Lernstandserhebungen (LSE):

1. *Wie viele Unterrichtsstunden haben Sie in dieser Klasse ungefähr für die unmittelbare Vorbereitung auf die LSE aufgewendet?* ca. _____ Unterrichtsstunden

2. *Welche der folgenden Optionen haben Sie in dieser Klasse als **unmittelbare Vorbereitung** auf die LSE genutzt? (Mehrfachnennungen möglich)* - Ich habe...

a)	…mit Testaufgaben früherer LSE üben lassen.	◯
b)	…alte Testaufgaben zur Verfügung gestellt.	◯
c)	…mit der Klasse gemeinsam die offizielle Internetseite des Ministeriums/des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) zu den LSE besucht.	◯
d)	…die Schüler auf die offizielle Internetseite des Ministeriums/des IQB hingewiesen.	◯
e)	…die Testsituation simuliert.	◯
f)	…in vorherigen regulären Klassenarbeiten zu den Aufgabenformaten der LSE ähnliche Aufgaben eingebaut.	◯
g)	...in den vorherigen regulären Klassenarbeiten Aufgaben aus vorherigen LSE eingebaut.	◯
h)	…im Unterricht die Beispielaufgaben von der offiziellen Homepage lösen lassen.	◯
i)	…alle Inhaltsbereiche vor den LSE noch einmal wiederholt.	◯
j)	…den Schülern empfohlen, noch einmal alle Inhaltsbereiche zu wiederholen.	◯
k)	…alle Prozessbereiche vor den LSE noch einmal wiederholt.	◯
l)	…den Schülern empfohlen, noch einmal alle Prozessbereiche zu wiederholen.	◯
n)	…eine Prozesskompetenz besonders üben lassen.	◯
o)	…keine dieser Optionen genutzt.	❑

Ich habe...

p) …eine hier nicht genannte Option gewählt, nämlich:_____

3. *Welche der folgenden Themen haben Sie im Unterricht angesprochen.* - Ich habe angesprochen...

		mehrfach	einmal	gar nicht
a)	…wie man die Aufgabenstellungen der LSE-Aufgaben richtig versteht.	☐	☐	☐
b)	…wie man den Aufgabenstellungen der LSE-Aufgaben die wichtigen Informationen entnimmt.	☐	☐	☐
c)	…wie man im Hinblick auf die speziellen Antwortformate der LSE richtig antwortet.	☐	☐	☐

4. *Von verschiedenen (Schul-)Buchverlagen gibt es für die drei Fächer Deutsch, Englisch und Mathematik so genannte „Vorbereitungshefte" für die LSE, die jedes Jahr neu aufgelegt werden und speziell auf die aktuell anstehenden LSE zugeschnitten sind.*

		in (fast) allen Stunden	in einigen Stunden	nein
a)	Haben Sie solche Hefte für die unmittelbare Vorbereitung genutzt?	☐	☐	☐
	Wenn Sie diese Vorbereitungshefte eingesetzt haben...	Alle	einige	nein
b)	…haben Sie solche Vorbereitungshefte genutzt, um einige oder alle Inhaltskompetenzen noch einmal zu wiederholen?	☐	☐	☐
c)	…haben Sie solche Vorbereitungshefte eingesetzt, um einige oder alle Prozesskompetenzen noch mal zu wiederholen?	☐	☐	☐
d)	…haben Sie Schüler darauf aufmerksam gemacht, dass es Vorbereitungshefte zu kaufen oder in der Schule zu leihen/ einzusehen gibt?	☐	☐	☐

5. *Weiter gibt es für das Fach Mathematik so genannte „Lernhilfen", Hefte mit Übungsaufgaben, die sich an den Kernlehrplänen orientieren.*

		in (fast) allen	in einigen	nein
a)	Haben Sie solche Hefte in den Unterrichtsstunden unmittelbar vor den LSE genutzt?	☐	☐	☐
	Wenn Sie diese Lernhilfen eingesetzt haben...	Alle	einige	nein
b)	…haben Sie solche Lernhilfen genutzt, um einige oder alle Inhaltskompetenzen noch einmal zu wiederholen?	☐	☐	☐
c)	…haben Sie solche Lernhilfen genutzt, um einige oder alle Prozesskompetenzen noch einmal zu wiederholen?	☐	☐	☐
d)	…haben Sie Schüler darauf aufmerksam gemacht, dass es Lernhilfen zu kaufen oder in der Schule zu leihen/einzusehen gibt?	☐	☐	☐

6. *Bitte geben Sie an, ob Sie folgende Strategien Ihren Schülern mit auf den Weg gegeben haben.* - Ich habe ihnen geraten...

		ja	nein
a)	…sich nicht zu lange mit einer Frage aufzuhalten.	☐	☐
b)	…sich mit dem Antwortformat vertraut zu machen.	☐	☐
c)	…alle Antworten in Betracht zu ziehen, bevor man sich entscheidet.	☐	☐
d)	…die Instruktionen und Fragen genau zu lesen.	☐	☐
e)	…bei Multiple-Choice-Fragen zu raten, wenn man die Antwort nicht weiß.	☐	☐
f)	…zuerst die Fragen zu beantworten, bei denen man sich sicher ist.	☐	☐
g)	…sich spontane Einfälle zu notieren.	☐	☐
h)	…auf die grammatikalische Einschränkung der möglichen Antworten zu achten.	☐	☐
i)	…im Zweifel die erste Idee zu wählen, weil dies meist das Beste ist.	☐	☐

Haben Sie zu einer anderen Strategie geraten?

k) ja, nämlich:_____

7. *Was denken Sie, wie wichtig sind die Lernstandserhebungen für…*

		gar nicht wichtig					sehr wichtig	weiß nicht
a)	…die Mehrheit Ihrer Schüler?	☐	☐	☐	☐	☐	☐	☐
b)	…die Mehrheit deren Eltern?	☐	☐	☐	☐	☐	☐	☐
c)	…die Schulleitung?	☐	☐	☐	☐	☐	☐	☐
d)	…die Mehrheit der anderen Deutsch-, Englisch- bzw. Mathematiklehrkräfte an Ihrer Schule, die aktuell mindestens eines dieser Fächer in einer achten Klasse unterrichten?	☐	☐	☐	☐	☐	☐	☐
e)	…Sie persönlich?	☐	☐	☐	☐	☐	☐	☐

Die nachfolgenden Fragen beziehen sich auf einen längeren Zeitraum:

8. *Wenn Sie sich noch einmal an die Klasse zurückerinnern, in der Sie zum ersten Mal die LSE durchgeführt haben, welche Aussage trifft dann zu? (Bitte nur eine Antwort in diesem Block ankreuzen.)*

a)	Ich bin zum ersten Mal als Lehrer mit LSE konfrontiert.	❑
b)	Ich bereite meine Schüler genauso intensiv auf die LSE vor wie vorher auch.	❑
c)	Ich bereite meine Schüler in diesem Jahr intensiver auf die LSE vor als frühere Jahrgänge.	❑
d)	Ich bereite meine Schüler in diesem Jahr weniger intensiv auf die LSE vor als frühere Jahrgänge.	❑

Wenn Sie in Frage 8 eine der Antwortmöglichkeiten b)-d) gewählt haben:

9. *Welche Gründe gibt es dafür?*

a)	Heute kenne ich mich mit den LSE besser aus.	☐
b)	In den vergangenen Jahren hat der Druck für mich *zugenommen*, bei den LSE gut abzuschneiden.	☐
c)	In den vergangenen Jahren hat der Druck für Schüler *zugenommen*, bei den LSE gut abzuschneiden.	☐
d)	In den vergangenen Jahren hat der Druck für mich *abgenommen*, bei den LSE gut abzuschneiden.	☐
e)	In den vergangenen Jahren hat der Druck für Schüler *abgenommen*, bei den LSE gut abzuschneiden.	☐

f) Andere Gründe, nämlich... _____

10. *Die nachfolgenden Fragen beziehen sich auf das ganze bisherige Schuljahr, nicht nur auf die unmittelbare Vorbereitung vor den LSE:*

		ja	nein
a)	Haben Sie eine Prozesskompetenz in dieser achten Klasse intensiver behandelt als in den Jahrgängen zuvor?	☐	☐
aa)	Falls ja, geschah dies mit Blick auf die aktuellen LSE?	☐	☐
b)	Haben Sie eine Prozesskompetenz in dieser achten Klasse weniger intensiv behandelt als in den Jahrgängen zuvor?	☐	☐
bb)	Falls ja, geschah dies mit Blick auf die aktuellen LSE?	☐	☐
c)	Gab es in der Fachgruppe Mathematik Absprachen, in diesem Schuljahr bestimmte Kompetenzen in der achten Klasse intensiver zu behandeln als in den Jahrgängen zuvor?	☐	☐

		ja	nein
cc)	Falls ja, geschah dies mit Blick auf die aktuellen LSE?	☐	☐
d)	Gab es in der Fachgruppe Mathematik Absprachen, in diesem Schuljahr bestimmte Kompetenzen in der achten Klasse weniger intensiver zu behandeln als in Jahrgängen davor?	☐	☐
dd)	Falls ja, geschah dies mit Blick auf die aktuellen LSE?	☐	☐
e)	Haben Sie einzelnen Schülern empfohlen, sich speziell auf die LSE vorzubereiten?	☐	☐
f)	Haben Sie Schülern empfohlen mit Blick auf die LSE Nachhilfe zu nehmen?	☐	☐
g)	Nehmen Schüler Ihrer Klasse nach Ihrem Kenntnisstand Nachhilfe speziell mit Blick auf die LSE?	☐	☐
h)	Haben Ihre Schüler nach Ihrem Kenntnisstand für die LSE außerhalb des Unterrichts besonders geübt?	☐	☐
i)	Haben Sie der Klasse empfohlen, sich speziell auf die LSE vorzubereiten?	☐	☐

Nun geht es um die Funktionen der LSE:

11.	*Wie stark stimmen Sie folgenden Aussagen zu bzw. lehnen Sie diese ab? Die LSE sind dazu gedacht…*	lehne völlig ab					stimme völlig zu
a)	…den Schülern eine Rückmeldung über ihre mathematischen Kompetenzen zu geben.	☐	☐	☐	☐	☐	☐
b)	…die Lehrer zu unterstützen, einfacher zu einer Zeugnisnote zu gelangen.	☐	☐	☐	☐	☐	☐
c)	…eine Rückmeldung über unser Schulsystem zu geben.	☐	☐	☐	☐	☐	☐
d)	…den Lehrkräften ein Arbeitszeugnis auszustellen.	☐	☐	☐	☐	☐	☐
e)	...für die Schüler eine Generalprobe für zentrale Prüfungen in der weiteren Schulzeit zu sein.	☐	☐	☐	☐	☐	☐
f)	…Lehrkräften Hinweise für die Klassenarbeits-konstruktion zu geben.	☐	☐	☐	☐	☐	☐
g)	…ein Schulranking zu erstellen.	☐	☐	☐	☐	☐	☐
h)	Lehrkräften Hinweise für die weitere Unterrichts-planung zu geben.	☐	☐	☐	☐	☐	☐
i)	…finanzielle und personelle Mittel an Schulen richtig zuzuweisen.	☐	☐	☐	☐	☐	☐

Als nächstes geht es um Ihre Überlegungen, auf die LSE vorzubereiten:

Wenn Sie Ihre Schüler auf die Lernstandserhebungen vorbereitet haben (sonst weiter mit Frage13):

12.		lehne völlig ab					stimme völlig zu
a)	Mir ist es wichtig, dass das Ansehen unserer Schule nicht unter schlechten Ergebnissen bei den LSE leidet.	☐	☐	☐	☐	☐	☐
b)	Ich möchte verhindern, dass der Schule finanzielle Mittel gestrichen werden.	☐	☐	☐	☐	☐	☐
c)	Ich möchte negative Rückmeldungen an mich verhindern, die aufgrund eines schlechten Abschneidens meiner Schüler geäußert werden könnten.	☐	☐	☐	☐	☐	☐
d)	Ich möchte nicht an einer Schule arbeiten, die bei den LSE schlecht abschneidet.	☐	☐	☐	☐	☐	☐

		lehne völlig ab					stimme völlig zu
e)	Ich möchte den Schülern ermöglichen, mit einem guten Ergebnis bei den LSE eine bessere Note zu bekommen.	☐	☐	☐	☐	☐	☐
f)	Ich möchte meinen Schülern die Chance geben, über die LSE Selbstvertrauen zu erlangen.	☐	☐	☐	☐	☐	☐
g)	Ich möchte den Erwartungen meiner Schüler genügen, sie optimal auf die LSE vorzubereiten.	☐	☐	☐	☐	☐	☐
h)	Ich möchte den Erwartungen der Schulleitung genügen, die Schüler optimal auf die LSE vorzubereiten.	☐	☐	☐	☐	☐	☐
i)	Ich möchte meinen Schülern helfen, ihren vollen Kompetenzumfang in den LSE zu zeigen.	☐	☐	☐	☐	☐	☐
j)	Ich möchte meinen Schülern die Angst vor den LSE nehmen.	☐	☐	☐	☐	☐	☐
k)	Ich möchte verhindern, dass Schüler lediglich deshalb eine falsche Antwort ankreuzen, weil sie das Aufgabenformat nicht richtig verstanden haben.	☐	☐	☐	☐	☐	☐
l)	Ich möchte den Erwartungen der Eltern genügen, die Schüler optimal auf die LSE vorzubereiten.	☐	☐	☐	☐	☐	☐

Wenn Sie Ihre Schüler auf die Lernstandserhebungen NICHT vorbereitet haben:

13.		lehne völlig ab					stimme völlig zu
a)	Ich möchte keine wertvolle Unterrichtszeit für die Vorbereitung auf die LSE verwenden.	☐	☐	☐	☐	☐	☐
b)	Ich möchte mich keiner ‚Testhysterie' unterwerfen.	☐	☐	☐	☐	☐	☐
c)	Ich möchte Instrumente wie die LSE nicht durch eine Vorbereitung legitimieren.	☐	☐	☐	☐	☐	☐
d)	Ich möchte mir keine Gedanken über eine spezielle Vorbereitung auf die LSE machen.	☐	☐	☐	☐	☐	☐
e)	Ich möchte verhindern, durch eine zu starke Vorbereitung auf die LSE ihre Aussagekraft der LSE zu gefährden.	☐	☐	☐	☐	☐	☐
f)	Ich möchte verhindern, dass Schülern durch eine Vorbereitung bei den LSE besser abschneiden als sie in Wirklichkeit sind.	☐	☐	☐	☐	☐	☐

TEIL 2: Ab jetzt geht es um Einstellungen zu anderen Elementen des Schulalltags und den Umgang mit den LSE sowie Einschätzungen der eigenen Person.

Zunächst geht es um Ihre Einstellung zu den Kernlehrplänen, zur Rechenschaftslegung und zur Einordnung Ihrer Schule zu einem der Standorttypen:

14.	*Bitte geben Sie den Grad der Zustimmung an.*	lehne völlig ab					stimme völlig zu
a)	Die Kernlehrpläne sind ein sinnvolles Instrument.	☐	☐	☐	☐	☐	☐
b)	Die Kernlehrpläne können helfen Unterricht zu gestalten.	☐	☐	☐	☐	☐	☐
c)	Die Kernlehrpläne sind wirkungsvoller als die vorherigen Lehrpläne.	☐	☐	☐	☐	☐	☐
d)	Die Kernlehrpläne tragen zu einem besseren Mathematikunterricht bei.	☐	☐	☐	☐	☐	☐

15.	*Wie stark stimmen Sie folgenden Aussagen über Rechenschaftslegung zu?*	lehne völlig ab					stimme völlig zu
a)	Die Ergebnisse zentraler Tests sollten veröffentlicht werden.	☐	☐	☐	☐	☐	☐
b)	Die Ergebnisse der LSE sollten zwingend nur schulintern bekannt werden.	☐	☐	☐	☐	☐	☐
c)	Die Ergebnisse der zentralen Abschlussprüfungen sollten zwingend nur schulintern bekannt werden.	☐	☐	☐	☐	☐	☐
d)	Die LSE sind kein geeignetes Instrument zur Rechenschaftslegung, weil sie fast nichts über die entscheidenden Sachverhalte in der Schule aussagen.	☐	☐	☐	☐	☐	☐

16.	*In den nächsten beiden Fragen geht es um den Standorttyp.*	Standorttyp 1 (Großstadt, viele Schüler mit Sprachproblemen..)	Standorttyp 2 (Kleinstadt, viele Schüler aus Akademikerfamilien…)	weiß nicht
	Welchem Standorttyp hat sich Ihre Schule für die LSE zugeordnet?	☐	☐	☐
		Ja	Nein	weiß nicht
	Halten Sie die Zuordnung für angemessen?	☐	☐	☐

Als nächstes möchten wir gern wissen, wie Sie Ihre Rolle im Mathematikunterricht und in der Schule insgesamt sehen. Einige der folgenden Fragen bzw. Aussagen mögen dabei sehr persönlich erscheinen. Wir möchten Sie dennoch bitten, diese Fragen nach Möglichkeit zu beantworten:

17.	*Die nachfolgenden Aussagen beziehen sich auf ihre aktuelle Situation als Lehrkraft. Wie stark würden Sie folgenden Aussagen zustimmen?*	lehne völlig ab					stimme völlig zu
a)	Wenn ich mir überlege, welche Herausforderungen mir als Mathematiklehrkraft begegnen, halte ich mich für einen guten Pädagogen.	☐	☐	☐	☐	☐	☐
b)	Ich erkenne oft nicht, ob ein Schüler etwas verstanden hat.	☐	☐	☐	☐	☐	☐
c)	Wenn ein Schüler eine Aufgabe falsch löst, ist mir der Grund dafür oft nicht klar.	☐	☐	☐	☐	☐	☐

		lehne völlig ab					stimme völlig zu
d)	Die Ursache für einen Schülerfehler sehe ich in der Regel sofort.	☐	☐	☐	☐	☐	☐
e)	Ich bin gut darin, Alltagsbezüge zur Mathematik herzustellen.	☐	☐	☐	☐	☐	☐
f)	Das für eine Klasse richtige Lerntempo treffe ich in der Regel genau.	☐	☐	☐	☐	☐	☐

18.	*Die nachfolgenden Aussagen beziehen sich auf das Erklären und Anregen im Mathematikunterricht.*	lehne völlig ab					stimme völlig zu
a)	Ich traue mir zu, Schüler anzuregen, selbst für komplizierte mathematische Sachverhalte und Probleme Erklärungen und Lösungswege zu finden.	☐	☐	☐	☐	☐	☐
b)	Ich bin mir sicher, mit meinem Unterricht auch bei insgesamt desinteressierten Schülern das Interesse für Mathematik wecken zu können.	☐	☐	☐	☐	☐	☐
c)	Ich traue mir zu, meinen Schülern auch komplizierte Beweise der Schulmathematik verständlich zu erklären.	☐	☐	☐	☐	☐	☐
d)	Ich weiß, dass ich Diskussionen über Lösungswege auch in „trägen Klassen“ anregen könnte, wenn ich das wollte.	☐	☐	☐	☐	☐	☐
e)	Ich weiß, dass es mir gelingen würde, anderen die Verbindungen zwischen zwei mathematischen Bereichen aufzuzeigen, auch wenn der Zusammenhang schwierig zu erkennen ist.	☐	☐	☐	☐	☐	☐
f)	Wenn ich es möchte, gelingt es mir nach wenigen Unterrichtsstunden, dass die Schüler in meinem Unterricht selbstständig Aufgaben bearbeiten und Probleme lösen.	☐	☐	☐	☐	☐	☐
g)	Ich kann einem Schüler auch schwer zu verstehende Bereiche der Mathematik erklären, sodass dieser es versteht.	☐	☐	☐	☐	☐	☐
h)	Meine Schüler zum Bearbeiten von mathematischen Aufgaben zu bewegen, bleibt für mich ein echtes Problem, wie sehr ich mich auch engagiere.	☐	☐	☐	☐	☐	☐

19.	*Wie stark treffen folgenden Aussagen über Ihre Bereitschaft und Möglichkeit zu Veränderungen an Ihrer Schule zu?*	trifft gar nicht zu					trifft völlig zu
a)	Ein möglicher Weg, damit sich an unserer Schule etwas bewegt, ist die Beteiligung an landes- und bundesweiten Wettbewerben.	☐	☐	☐	☐	☐	☐
b)	Es erscheint mir sinnvoll, über Veränderungen in der Schule regelmäßig gemeinsam nachzudenken.	☐	☐	☐	☐	☐	☐
c)	Ich bin an Veränderungsprozessen meistens mehr beteiligt als viele andere Kollegen.	☐	☐	☐	☐	☐	☐
d)	Oft gehen Veränderungsprozesse mehr von mir aus als von anderen.	☐	☐	☐	☐	☐	☐

20.	*In den folgenden Aussagen geht es um Arbeit als Pflicht. Wie sehr stimmen Sie diesen Aussagen zu?*	lehne völlig ab					stimme völlig zu
a)	Ich bin immer vorbereitet.	☐	☐	☐	☐	☐	☐
b)	Ich achte auch auf Details.	☐	☐	☐	☐	☐	☐
c)	Meine Hausarbeit mache ich sofort.	☐	☐	☐	☐	☐	☐
d)	Meine Pläne setze ich in die Tat um.	☐	☐	☐	☐	☐	☐
e)	Ich mache Pläne und halte mich daran.	☐	☐	☐	☐	☐	☐
f)	Ich benötige oft mehr Zeit als nötig.	☐	☐	☐	☐	☐	☐
g)	Mir fällt es schwer, mich an die Arbeit zu setzen.	☐	☐	☐	☐	☐	☐
h)	Ich mache gerade so viel wie nötig.	☐	☐	☐	☐	☐	☐
i)	Ich überprüfe meine Arbeit.	☐	☐	☐	☐	☐	☐
j)	Wenn möglich, weiche ich Pflichten aus.	☐	☐	☐	☐	☐	☐

Weitere interessiert uns, wie groß das Interesse an den LSE aus Ihrer Sicht ist:

21.	*Bitte geben Sie jeweils an, wie sehr die folgenden Aussagen zutreffen.*	trifft gar nicht zu					trifft völlig zu
a)	Mich interessiert, inwiefern die LSE bislang fehlende Informationen liefern.	☐	☐	☐	☐	☐	☐
b)	Ich möchte gerne meine Schüler in die Auseinandersetzung mit den LSE-Ergebnissen einbeziehen, um daraus zu lernen.	☐	☐	☐	☐	☐	☐
c)	Ich bin gerne bereit, meine Kollegen bei der Beschäftigung mit den LSE-Ergebnissen zu unterstützen.	☐	☐	☐	☐	☐	☐
d)	Ich mache mir keine Gedanken über die LSE.	☐	☐	☐	☐	☐	☐
e)	Mich interessiert, welche Veränderungen andere Lehrer aus den LSE-Ergebnissen für ihren Unterricht abgeleitet haben.	☐	☐	☐	☐	☐	☐

22.	*Wie stark treffen folgenden Aussagen über das Interesse der Eltern und Schüler in Bezug auf die LSE zu?*	trifft gar nicht zu					trifft völlig zu
a)	Die Eltern der Schüler an unserer Schule interessieren sich für das Abschneiden ihrer Kinder bei den LSE.	☐	☐	☐	☐	☐	☐
b)	Die Eltern unserer Schule interessieren sich dafür, wie die Klasse ihres Kindes abgeschnitten hat.	☐	☐	☐	☐	☐	☐
c)	Die Eltern unserer Schule interessieren sich dafür, wie die Schule insgesamt bei den LSE abgeschnitten hat.	☐	☐	☐	☐	☐	☐
d)	Meine Schüler interessiert, wie unsere Schule in der Vergangenheit bei den LSE abgeschnitten hat.	☐	☐	☐	☐	☐	☐

23.	*Wie stark treffen folgenden Aussagen über das Verhalten der Schulleitung in Bezug auf die LSE zu?*	trifft gar nicht zu					trifft völlig zu
a)	Der Schulleitung ist es sehr wichtig, dass die Schule gut bei den LSE abschneidet.	☐	☐	☐	☐	☐	☐
b)	Die Schulleitung interessiert sich sehr für die LSE.	☐	☐	☐	☐	☐	☐

	Wenn eine Klasse in einem Fach bei den LSE…	trifft gar nicht zu					trifft völlig zu
c)	… unterdurchschnittlich abschneidet, fragt die Schulleitung die unterrichtende Fachlehrkraft, welche Erklärung sie für das schlechte Abschneiden hat.	☐	☐	☐	☐	☐	☐
d)	…gut abschneidet, hebt die Schulleitung die unterrichtende Fachlehrkraft besonders hervor.	☐	☐	☐	☐	☐	☐
	Wenn die Klassen einer Lehrkraft mehrfach schlechtere Ergebnisse bei den LSE erreichen als andere Klassen an unserer Schule…						
e)	…bittet die Schulleitung die Lehrkraft zum Gespräch.	☐	☐	☐	☐	☐	☐
f)	…bietet die Schulleitung Unterstützung an.	☐	☐	☐	☐	☐	☐
g)	…stellt die Schulleitung lediglich Forderungen, ohne jedoch Hilfe anzubieten.	☐	☐	☐	☐	☐	☐

24.	*Bitte kreuzen Sie an, wie die Ergebnisse der LSE in der Öffentlichkeit wahrgenommen werden.*	ja	nein
a)	Ich weiß, dass es regelmäßig Anfragen der regionalen Medien an unsere Schule zu den Ergebnissen der LSE gibt.	☐	☐
b)	Die regionalen Medien berichten über das Abschneiden der Schulen in dieser Stadt.	☐	☐

Zuletzt wüssten wir gern, wie Sie sowie Ihre Kollegen und das Ministerium mit den Ergebnissen in der Vergangenheit umgegangen sind:

25.	*Wie stark treffen folgenden Aussagen über den Umgang mit den Ergebnissen der LSE zu?*	trifft gar nicht zu					trifft völlig zu
a)	Ich denke, dass die Beschäftigung mit den Rückmeldungen aus LSE wichtig ist.	☐	☐	☐	☐	☐	☐
b)	Ich habe mich angestrengt, die Ergebnisse zu analysieren.	☐	☐	☐	☐	☐	☐
c)	Ich habe mich bemüht, die Ergebnisse für meinen Unterricht zu verwenden.	☐	☐	☐	☐	☐	☐
d)	Ich habe bisher nicht sehr viel Energie in die Beschäftigung mit den Fähigkeitsniveau-Verteilungen gesteckt.	☐	☐	☐	☐	☐	☐
e)	Die Auseinandersetzung mit den LSE-Ergebnissen war nützlich für meinen Unterricht.	☐	☐	☐	☐	☐	☐

26.	*Wie stark treffen folgenden Aussagen über ein mögliches Schulranking zu?*	trifft gar nicht zu					trifft völlig zu
a)	Ich denke, dass die bei den LSE erhobenen Leistungsdaten beim Ministerium intern für ein Schulranking genutzt werden.	☐	☐	☐	☐	☐	☐
b)	Ich denke, dass die Daten der LSE in der Zukunft für ein öffentliches Schulranking genutzt werden.	☐	☐	☐	☐	☐	☐

27. *Wie stark treffen folgenden Aussagen über die erfahrene Unterstützung in Bezug auf die LSE zu?*

	Wenn die Klassen einer Lehrkraft mehrfach schlechtere Ergebnisse bei den LSE erreichen als andere Klassen an unserer Schule...	trifft gar nicht zu					trifft völlig zu
a)	…diskutieren wir in der Fachgruppe gemeinsam über Ursachen.	☐	☐	☐	☐	☐	☐
b)	…suchen wir in der Fachgruppe gemeinsam Wege zur Verbesserung.	☐	☐	☐	☐	☐	☐
c)	…wird die betroffene Lehrkraft von ihren Kollegen mit dem Problem der schlechten Ergebnisse allein gelassen.	☐	☐	☐	☐	☐	☐

28. *Inwieweit haben die Ergebnisse der LSE Veränderungen herbeigeführt?*

	individuelle Veränderungen Das Abschneiden unserer Schüler bei…	trifft gar nicht zu					trifft völlig zu
a)	…den LSE gab mir schon **mehrfach** Anlass, Veränderungen im Mathematik-Unterricht zu überlegen.	☐	☐	☐	☐	☐	☐
b)	…der **letzten LSE,** an denen ich selbst beteiligt war, gab mir Anlass, Veränderungen im Mathematik-unterricht zu überlegen.	☐	☐	☐	☐	☐	☐
c)	… den LSE gab mir schon **mehrfach** Anlass, mich mit bestimmten mathematischen Teilgebieten intensiver zu beschäftigen.	☐	☐	☐	☐	☐	☐
d)	…der **letzten LSE**, an denen ich selbst beteiligt war, gab mir Anlass, mich mit bestimmten mathematischen Teilgebieten intensiver zu beschäftigen.	☐	☐	☐	☐	☐	☐
	Veränderungen durch die Fachgruppe Aufgrund der Ergebnisse…						
e)	…unserer Schüler haben wir innerhalb der Fachgruppe **mehrfach** Maßnahmen zur besseren Förderung der Schüler diskutiert.	☐	☐	☐	☐	☐	☐
f)	…bei den **letzten LSE** haben wir innerhalb der Fachgruppe Maßnahmen zur besseren Förderung der Schüler diskutiert.	☐	☐	☐	☐	☐	☐
g)	…unserer Schüler haben wir innerhalb der Fachgruppe **mehrfach** konkrete Maßnahmen zur Verbesserung des Unterrichts diskutiert.	☐	☐	☐	☐	☐	☐
h)	…bei den **letzten LSE** haben wir innerhalb der Fachgruppe konkrete Maßnahmen zur Verbesserung des Unterrichts diskutiert.	☐	☐	☐	☐	☐	☐

Haben Sie noch Anmerkungen zum Fragebogen oder den LSE, dann können Sie die hier notieren:

_____

_____

_____

_____

Vielen herzlichen Dank für Ihre Teilnahme!

R-Code

```
## Bootstrap Version1

bootstrap.poLCA <- function(NameLCA, KlassenAnz, VariablenAnz, Faelle, BootstrapAnz, Niveau){

	BootLCA_Chi <- numeric(BootstrapAnz)

	BootLCA_Log <- numeric(BootstrapAnz)

	BootLCA_Chi_x <- numeric(BootstrapAnz)

	Niveau_Wert <- Niveau*BootstrapAnz

	t <- 1

	for (t in 1:BootstrapAnz){

		Datensatz <- poLCA.simdata(N=Faelle, probs=NameLCA$probs, nclass=KlassenAnz,
P=NameLCA$P, missval=TRUE, pctmiss=0.003)

		LCA_sim <-poLCA(f, Datensatz$dat, nclass=KlassenAnz, na.rm=FALSE)

		BootLCA_Chi_x[t] <- t

		BootLCA_Chi[t] <- LCA_sim$Chisq

		BootLCA_Log[t] <- LCA_sim$Gsq

	}

	BootLCA_Chisort <- sort(BootLCA_Chi)

	Chi_Hilfeswert <- which(BootLCA_Chisort <= NameLCA$Chisq)

	Chi_p_Wert <- 1-(max(Chi_Hilfeswert)/BootstrapAnz)

	BootLCA_Logsort <- sort(BootLCA_Log)

	Log_Hilfeswert <- which(BootLCA_Logsort <= NameLCA$Gsq)

	Log_p_Wert <- 1-(max(Log_Hilfeswert)/BootstrapAnz)

	Ausgabe <- c(NameLCA$Chisq, BootLCA_Chisort[Niveau_Wert], Chi_p_Wert, NameLCA$Gsq,
BootLCA_Logsort[Niveau_Wert], Log_p_Wert)

	print(Ausgabe)

	plot(BootLCA_Chi_x,BootLCA_Chisort)

}
```

Eidesstattliche Erklärung

Hiermit erkläre ich, Jörn Sebastian Hahn, geb. am 05.09.1983, an Eides statt gemäß der Promotionsordnung §7(2) des Fachbereichs für Bildungswissenschaften:

dass ich die eingereichte Dissertation selbständig verfasst habe,

dass ich in vorausgegangene Promotionsverfahren in dem betreffenden Fach oder in einem anderen Fach nicht endgültig gescheitert bin,

dass ich bei der Abfassung der Dissertation nur die angegebenen Hilfsmittel benutzt und alle wörtlich oder inhaltlich übernommenen Stellen als solche gekennzeichnet habe

und dass ich die Dissertation nur in diesem Promotionsverfahren eingereicht habe.

Bremen, der 07.04.2014 _____

Lebenslauf

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.